Habitatmodellierung europäischer Waldgesellschaften mittels Maximum-Entropie-Methode und Genetischem Programmieren

Diplomarbeit

zur Erlangung des akademischen Grades
Diplom-Bioinformatiker

Friedrich-Schiller Universität Jena

Fakultät für Mathematik und Informatik
und
Biologisch-Pharmazeutische Fakultät

vorgelegt von
Dennis Görlich
geboren am 02. Juni 1983 in Hagen in Westf.

1. Betreuer: PD Dr. Gottfried Jetschke

2. Betreuer: PD Dr. Peter Dittrich

Jena, den 31. Juli 2008

Zusammenfassung

In dieser Arbeit werden Modelle für die Verbreitung von Eichen-Hainbuchen-Wäldern in Europa erstellt und auf ein Klimawandel-Szenario angewendet. Dazu werden zwei Methoden verwendet, die Maximum-Entropie-Methode (MaxEnt) und Genetisches Programmieren (GP). Die Modelle basieren auf Klimadaten, die frei auf worldclim.org zur Verfügung stehen und Vegetationsdaten aus der Karte der natürlichen potentiellen Vegetation Europas von Bohn. Die Arbeit zeigt, dass MaxEnt schon bei kleinen Stichprobengrößen gute Modellergebnisse erzeugt und dass sich die Modellgüte in Abhängigkeit von negativ korrelierten Variablen verbessert. Die Variablenauswahl durch eine vorherige Diskriminanzanalyse und die schrittweise Erstellung des Modells anhand dieser Auswahl zeigen, wie die Modellgüte mit der Variablenanzahl zunimmt.

Die Modellerstellung mittels GP auf Testproblemen hat ergeben, dass GP in der Lage ist Funktionen zu lernen, die den Zusammenhang zwischen Habitat und Umwelt beschreiben. Allerdings kann das GP den großen Suchraum nicht effektiv durchsuchen, so dass für praktische Probleme keine brauchbaren Lösungen gefunden wurden.

Die Anwendung der MaxEnt-Modelle auf Umweltdaten aus dem IPCC-Klimaszenario A2a für das Jahr 2080 zeigt, wie sich die Bedingungen für Eichen-Hainbuchen-Wälder verändern und wo diese Waldgesellschaften in der Zukunft optimale Klimabedingungen vorfinden könnten. Alle Projektionen zeigen einen „Osttrend“. Dieser Trend manifestiert sich in der Verringerung der Habitatgüte am Westrand der heutigen Verteilung und durch eine Erweiterung des Habitats im Osten Europas.

Inhaltsverzeichnis

1 Einleitung **9**

2 Habitatmodellierung **11**

2.1 Einordnung von Habitatmodellen in allgemeine Modellkategorien . . . 11

2.2 Prinzipien der Modellbildung 13

2.2.1 Konzeptionelles Modell 13

2.2.2 Statistische Modellformulierung 16

2.2.3 Kalibrierung 18

2.2.4 Vorhersage 18

2.2.5 Modellevaluation 19

2.3 Die Maximum-Entropie-Methode 25

2.3.1 Maximum-Entropie-Habitatmodelle 25

2.3.2 Interpretation der Maxent-Verteilung 26

2.4 Genetisches Programmieren zur Habitatmodellierung 28

2.4.1 Algorithmus 28

2.4.2 Repräsentation der Individuen 29

2.4.3 Funktionsmenge 30

2.4.4 Charakterisierung des Suchraums 31

2.4.5 Genotyp-Phänotyp-Abbildung 34

2.4.6 Zielfunktion 35

2.4.7 Selektion 36

2.4.8 Variation 37

2.4.9 Implementierungsdetails 38

3 Eichen-Hainbuchen-Wälder **45**

3.1 Ökologische Motivation 45

3.2 Konkrete Modellplanung 47

3.2.1 Konzeptionelles Modell 47

3.2.2 Datengrundlage 49

4 Anwendung der Maximum-Entropie-Methode **55**

4.1 Evaluierung der Maximum-Entropie-Methode in verschiedenen Szenarien . . . 56
4.1.1 Einfluss der Stichprobengröße . . . 56
4.1.2 Einfluss korrelierter Variablen . . . 59
4.1.3 Variablenselektion nach Diskriminanzanalyse . . . 60
4.1.4 Zusammenhang von AUC und R^2 . . . 68
4.2 Modelle europäischer Eichen-Hainbuchen-Wälder . . . 69
4.2.1 Europäische Eichen-Hainbuchen-Wälder . . . 69
4.2.2 Mitteleuropäische Eichen-Hainbuchen-Wälder . . . 72
4.2.3 Französisch-Süddeutsche Eichen-Hainbuchen-Wälder . . . 74
4.2.4 Hercynisch-Polonische Eichen-Hainbuchen-Wälder . . . 76
4.2.5 Slovakische Eichen-Hainbuchen-Wälder . . . 78
4.3 Diskussion . . . 79

5 Anwendung des Genetischen Programmierens 83
5.1 Künstliche Testdaten . . . 83
5.2 Eichen-Hainbuchen-Daten . . . 89
5.3 Diskussion . . . 89

6 Abschließende Diskussion und Ausblick 91

Literaturverzeichnis 94

Anlagen 101

Abkürzungen

AUC Area Under Curve, die Fläche unter einem ROC-Graphen

FFH Fauna-Flora-Habitat-Richtlinie

GP Genetisches Programmieren

GIS Geographisches Informationssystem oder general iterative scaling

GNU GNU is not Unix: Projekt für lizensfreie Software

GSL GNU Scientific Library

HabitatGP Genetisches Programm zur Habitatmodellierung

LGP Lineares genetisches Programmieren

MaxEnt Maximum-Entropie-Ansatz zur Habitatmodellierung

MAXENT Maximum-Entropie-Software

ME Maximum-Entropie-Methode zur Schätzung von Wahrscheinlichkeitsverteilungen

ME-Prinzip Maximum-Entropie-Prinzip

PE-Ratio Verhältnis von vorhergesagter zu erwarteter Fläche(predicted-expected-ratio)

PNV Potentielle Natürliche Vegetation

ROC Receiver-Operator-Curve

1 Einleitung

Diese Arbeit beschäftigt sich mit der Erstellung von Habitatmodellen für europäische Waldgesellschaften. Das Ziel ist es, die Maximum-Entropie-Methode und Genetisches Programmieren zur Modellierung von Eichen-Hainbuchen-Wäldern zu verwenden. Habitatmodelle werden immer häufiger in der Ökologie angewandt, um in verschiedenen Fragestellungen Entscheidungshilfen zu geben. Neben dem reinen Erkenntnisgewinn in Bezug auf die Ansprüche einer Art an ihre Umwelt und der Beschreibung eines Habitats finden Habitatmodelle Anwendung im Artenschutz und Artenmanagment [31, 13, 51].

Ein Großteil der Arbeiten im Bereich der Habitatmodellierung beschäftigt sich mit Modellen für Tierarten und bedrohte Pflanzenarten. Der Anteil an Arbeiten zu Baumarten ist gering und beschränkt sich auf wenige lokale oder nationale Studien. Zum Beispiel hat Kölling für Deutschland [32] und speziell für Bayern [33] mittels Klimahüllen (climatic envelopes) Modelle für einzelne Baumarten (Fichte und Buche) erstellt und diese auch in einem Klimawandelszenario betrachtet. Die vorliegende Arbeit ist eine der ersten, die Habitatmodelle von Waldgesellschaften für ganz Europa betrachtet. Allerdings übersteigt die Erstellung von Modellen für alle in Europa vorkommenden Waldgesellschaften den Umfang einer Diplomarbeit. Daher werden hier beispielhaft Eichen-Hainbuchen-Wälder (Verband Carpinion betuli) behandelt. Diese Modelle werden anschließend auf Umweltdaten für 2080 angewandt, um eine Vorstellung davon zu bekommen, wie das Areal sich verändert. In dieser Arbeit wird eine für die Habitatmodellierung relativ neue, aber schon erfolgreich eingesetzte, Methode verwendet, die Maximum-Entropie-Methode (MaxEnt, [49]). Maximum-Entropie ist ein allgemeiner Lösungsansatz zur Schätzung unbekannter Wahrscheinlichkeitsverteilungen [30] und wurde erfolgreich in der Physik und dem Maschinenlernen eingesetzt. Zur Berechnung von Modellen mittels MaxEnt steht eine frei verfügbare Software (MAXENT) auf der Seite der Autoren zum Download bereit.[1]

Außerdem wird in dieser Arbeit genetisches Programmieren (GP) verwendet, um Habitatmodelle zu berechnen. GP ist in der Habitatmodellierung kein neuer An-

[1] (http://www.cs.princeton.edu/ schapire/maxent/)

satz. Bisher wurde GP verwendet, um Regelsätze zu lernen, die die Verteilung einer Art in Abhängigkeit der Umwelt zu beschreiben [54, 38]. Im Gegensatz dazu soll das hier vorgestellte GP eine geschlossene Funktion lernen, die die Umwelt als Eingabe nimmt und die Güte eines potentiellen Habitats berechnet.
Die Arbeit weist folgende Struktur auf. In Kapitel 2 wird eine allgemeine Einführung in die Aufgabenstellung der Habitatmodellierung gegeben. Das Kapitel beinhaltet Vorüberlegungen und Schritte zur Erstellung eines Habitatmodells und eine kurze Übersicht über mögliche Gütemaße, die zur Bewertung der Modelle herangezogen werden. Außerdem werden die beiden verwendeten Modellierungstechniken erläutert. In Kapitel 3 wird das Fallbeispiel Eichen-Hainbuchen-Wälder vorgestellt und die Modellplanung konkretisiert. Die verwendeten Vegetations- und Umweltdaten werden beschrieben. Die Ergebnisse der Modellierung mittels der Maximum-Entropie-Software (MAXENT) werden in Kapitel 4 dargestellt. Es wird eine Beschreibung der gelernten Verteilungen, wichtiger Umweltvariablen und der Projektionen auf Umweltdaten für 2080 für die einzelnen Teilmodelle gegeben. Die Ergebnisse, die das genetische Programmieren mit Testdaten erzielt hat, und aufgetretene Probleme, werden in Kapitel 5 erörtert. In Kapitel 6 werden die Ergebnisse rekapituliert und ein Ausblick auf weitere Anwendungsmöglichkeiten und Verbesserungen gegeben.

2 Habitatmodellierung

In der Natur kann beobachtet werden, dass Arten ein bestimmtes Gebiet besiedeln, dort leben, Nahrung suchen und sich fortpflanzen. Dieses räumliche Gebiet wird als **Habitat** der Art bezeichnet. In der Vegetationsökologie wird lokal vom **Standort** bzw. global vom **Areal** einer Art oder Pflanzengesellschaft gesprochen.
Warum eine Art in manchen Gebieten zu finden ist und in anderen nicht, lässt sich durch die Ansprüche der Art an ihre Umwelt erklären, die ökologische Nische der Art. Eine Art kann in einem Gebiet nur überleben, wenn diese Ansprüche erfüllt werden.
Ein Modell stellt eine Abbildung dar, durch die versucht wird, die Wirklichkeit vereinfacht darzustellen. Die Zielstellung des Modells und der Grad der Abstraktion bestimmen, welche Merkmale der Realität das Modell wiedergeben soll. Das Modell kann dann verwendet werden, um zu neuen Erkenntnissen über die modellierten Zusammenhänge zu gelangen, oder Prognosen zu erstellen.
Mit Habitatmodellen wird die Beziehung zwischen dem Standort einer Art oder Gesellschaft und ausgewählten standortsbestimmenden Einflussgrößen hergestellt.

2.1 Einordnung von Habitatmodellen in allgemeine Modellkategorien

Modelle können nach verschiedenen Kriterien charakterisiert werden. Eine erste Beschreibung wird durch die zeitliche Natur gegeben. Wird ein Prozess modelliert, spricht man von **dynamischen**, bei einer Erfassungen des Zustands eines Systems von **statischen** Modellen. Ein Modell ist **stochastisch**, wenn Zufallskomponenten in der Auswertung vorkommen. Insbesondere erzeugen mehrere Auswertungen des Modells unterschiedliche Ergebnisse. **Deterministische** Modelle zeigen hingegen bei jeder Auswertung exakt dasselbe Ergebnis.
Nach Levins [34] können weitere Modelleigenschaften in Bezug auf Genauigkeit, Generalität und Realismus definiert werden, wobei nur zwei dieser drei Kriterien in

einem Modell zu Genüge erfüllt werden können. Diese Kategorisierung von Levins wird kritisch diskutiert [45, 35, 40, 44, 41]. Orzack behauptet, dass Levins mit dieser Kategorisierung Modelle willkürlich einteilt. Odenbaugh argumentiert, dass diese Kategorisierung sich auf die praktische Anwendung bezieht und dort durchaus ihre Gültigkeit behält. Levins' Konzept bietet aber trotzdem die Möglichkeit, Modelle, in Bezug auf das Ziel der Modellierung einzugrenzen (vgl. Guisan und Zimmermann [21]). Aus Levins Sichtweise ergeben sich drei weitere Modellkategorien: empirische, mechanistische und analytische Modelle (siehe Abbildung 2.1).
Alle vorgestellten Kategorisierungen sind uneingeschränkt miteinander vereinbar, so dass ein Modell zum Beispiel stochastisch, dynamisch und analytisch sein kann.

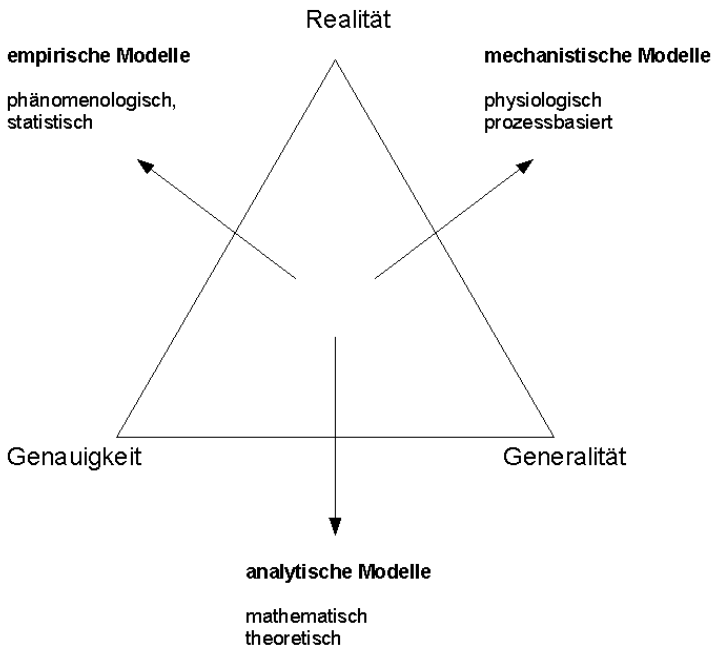


Abbildung 2.1: Charakterisierung von Modellen in Bezug auf Realität, Generalität und Genauigkeit (verändert nach Guisan und Zimmermann [21], dort zitiert aus Levins [34])

Habitatmodelle sind in der Regel empirisch, da sie aus beobachteten Daten erstellt werden. Sie sind deterministisch, da ihrer Berechnung ein fester Zusammenhang von Habitateigenschaften und Habitatgüte zu Grunde liegt, und sie sind statisch, da sie einen Gleichgewichtszustand erfassen sollen [21].
Eine treffende Definition wird im Geo-Informatik-Lexikon der Universität Rostock gegeben: „Ein Habitatmodell ist eine flächenhafte, semiquantitative Beschreibung eines Landschaftsausschnittes als potenzieller Lebensraum von Pflanzen und Tieren auf der Grundlage von strukturellen und funktionalen Abhängigkeiten zwischen wesentlichen Habitateinflussgrößen." [1]

2.2 Prinzipien der Modellbildung

Die Erstellung von Habitatmodellen erfolgt in 5 Schritten (nach Guisan und Zimmermann [21]):

1. konzeptionelles Modell,
2. statistisches Modell,
3. Kalibrierung,
4. Vorhersage,
5. Evaluierung.

Die Einteilung in einen konzeptionellen und einen statistischen Modellteil ist eine Möglichkeit die Modellbildung in Phasen einzuteilen und zu strukturieren. Im Vergleich dazu unterteilt Austin [5] das konzeptionelle Modell in ein ökologisches Modell und eine Datenmodell und trennt dadurch die theoretischen Überlegungen stärker von den mehr technisch geprägten Eigenschaften der Daten ab.

2.2.1 Konzeptionelles Modell

Das konzeptionelle Modell umfasst eine Reihe von Annahmen und Festlegungen, die sich auf das vorhandene Vorwissen und die Zielstellung des Modells beziehen.
In der Habitatmodellierung wird für jedes Modell angenommen, dass sich die zu modellierende Art in einem Gleichgewichtszustand in Bezug auf ihr Habitat und die Umwelt befindet [21]. Dies hat zur Folge, dass Habitatmodelle immer statische Modelle sind. Dieser Gleichgewichtszustand ist in vielen Fällen allerdings nur ein **Pseudogleichgewicht**, da sich viele Pflanzenarten zum Beispiel noch in ihrer nacheiszeitlichen Ausbreitung befinden [20]. Das Modell stellt in diesen Fällen eine Momentaufnahme dar und kann den wirklichen Zusammenhang zwischen Habitat und Umwelt nur unvollständig wiedergeben [21]. Wird die Gleichgewichtsannahme fallen gelassen, kann das Modell nur dynamisch und stochastisch sein [21], und es werden zusätzlich Informationen über das Verhalten der Art oder Gesellschaft bei Änderungen in den Prädiktorvariablen benötigt, so dass diese Form der Habitatmodellierung nur für gut untersuchte Arten sinnvoll ist [21, 52].
Weitere Punkte, die in dieser Phase der Modellbildung berücksichtigt werden sollten sind die Abgrenzung des **Untersuchungsgebiets** auf einer lokalen, regionalen oder globalen Skala, die **räumliche Auflösung** der verfügbaren Daten, die **Va-**

riablenselektion, die Methodik der **Stichprobenerhebung** (sampling design), die Entscheidung für die Modellierung von **Arten oder Gesellschaften** und die Berücksichtigung weiterer bekannter Informationen. Das konzeptionelle Modell legt somit den Rahmen für das Vorgehen im Modellierungsprozess fest.

Untersuchungsgebiet

Soll die aktuelle Verteilung einer Art durch ein Modell erfasst werden, so wird eine Stichprobe von Orten benötigt, an denen die Art vorkommt und Orten, an denen die Art nicht vorkommt. Diese Stichprobe wird in einem bestimmten, vorher festgelegtem Untersuchungsgebiet erhoben. Ein Modell, das mit dieser Stichprobe kalibriert wird, ist zunächst nur für dieses Untersuchungsgebiet gültig. Ob das Modell auch auf andere Gebiete angewandt werden kann, muss durch theoretische Überlegungen validiert werden. Dies kann zum Beispiel davon abhängen, ob für das neue Gebiet alle ins Modell eingegangenen Umweltvariablen zur Verfügung stehen. Die Größe des Untersuchungsgebiets (lokales, regionales oder globales Modell) kann Einfluss auf die weiteren Entscheidungen im konzeptionellen Modell haben.

Variablenselektion

Werden Habitatmodelle aus Feldbeobachtungen erstellt, so ist es vor der Stichprobenerhebung notwendig eine Menge von **Prädiktorvariablen** auszuwählen, die zur Modellierung verwendet werden. Diese Prädiktorvariablen werden dann für jeden Stichprobenpunkt bestimmt. Sind die Werte der Prädiktorvariablen an den Stichprobenpunkten bekannt, zum Beispiel globale digitale Klimadaten, die für alle Punkte schon definiert sind, so kann die Auswahl geeigneter Variablen auch nach der Stichprobenerhebung erfolgen.
In jedem Fall sollten die Variablen passend für die Zielstellung ausgewählt werden. So spielen zum Beispiel Boden und Geologie für Pflanzen auf der lokalen Skala eine große Rolle, während in einem globalen Modell klimatische Bedingung die primären Umweltfaktoren sind. Wichtig ist auch der Typ der Variablen. Es kann zwischen indirekten Variablen, direkten Variablen und Ressource unterschieden werden [21]. Je nach Zielstellung des Modells sind Variablen eines oder mehrerer Typen zur Modellerstellung sinnvoll (siehe auch Abbildung 3.4 auf Seite 50).

Arten versus Gesellschaften

Ein weiterer Teil der konzeptionellen Modellformulierung betrifft die Ausrichtung des Modells in Bezug auf Arten oder Gesellschaften. Üblicherweise wird das Habitat einer einzelnen Art modelliert. Erweiterte Ansätze erlauben es, mehrere Arten gleichzeitig zu modellieren (Community-Methoden) und so zum Beispiel Interaktionen zwischen den Arten zu erfassen. Die Modellierung von Waldgesellschaften in dieser Arbeit ist nicht im Sinne dieser Community-Methoden zu verstehen, da die Gesellschaft hier als Einheit und nicht als Menge von Arten aufgefasst wird (siehe auch Abschnitt 3.2.1).

Betrachtung zur Nische

Da Habitatmodelle die Verteilung einer Art als Funktion der Umweltparameter beschreiben, ist es unumgänglich, dass das Modell die Nische dieser Art erfasst und mitmodelliert. Verschiedene Techniken basieren auf unterschiedlichen Nischenkonzepten, welche sich aus den verschiedenen Theorien zur ökologischen Nische ableiten lassen. Die ersten Theorien von Grinnell und Elton, Anfang des 20. Jahrhunderts, basierten auf der verbalen Beschreibungen von Ansprüchen und Lebensräumen von Arten (vgl. Schöner in [10]). Diese bieten kaum Ansatzpunkte für mathematische Modelle. Hutchinson hat in den 1950er Jahren ein formales Nischenkonzept erstellt, welches die Umwelt als Achsen eines n-dimensionalen Raumes definiert und die Nische einer Art als Hypervolumen in diesem Raum. Im einfachsten Fall lässt sich dieses durch unabhängige Intervalle für jede Variable beschreiben, die einen Hyperquader aufspannen [29].
In den 1960er Jahren setzte sich das Utilisierungskonzept durch. Dem n-dimensionalen Raum der Umweltfaktoren wird eine weitere Achse hinzugefügt, die die Stärke der Nutzung der Nische beschreibt. Die Nischennutzung einer Art in Bezug auf eine einzelne Achse (**Gradient**) wird als **Response** der Art bezeichnet. Die Response kann, je nach Modellaurichtung, als Vorkommenswahrscheinlichkeit, Reproduktionsrate, Populationsdichte oder ähnliches interpretiert werden. Welche Formen eine Responsekurve annehmen kann, ist bis heute Forschungsgegenstand und Diskussionsthema (siehe z.B. [28, 42])).
Stehen Arten zueinander in Konkurrenz um Ressourcen, drückt sich dies im Raum der Umweltvariablen durch eine Überschneidung ihrer Nischen aus. Diese Überschneidung führt dazu, dass beide Arten diese Ressource nicht mehr optimal nutzen können und die Response sich im Überlappungsgebiet verringert. In diesem Fall spricht man von der **realisierten Nische** der Art. Im Gegensatz dazu besitzt eine

Art ohne Konkurrenz einen größere Nische, die **fundamentale Nische**. Ein Habitatmodell erfasst die realisierte Nische [21], da die Daten meist aus Verteilungen erhoben werden, die sich unter Konkurrenz ausgebildet haben.

Stichprobenerhebung

In der Regel wird die Stichprobe, aus der das Modell gelernt wird, erst erhoben, wenn die konzeptionelle Modellformulierung abgeschlossen ist, da die Art der Response, die zu erhebenden Umweltvariablen und das Untersuchungsgebiet feststehen. In einigen Fällen werden Modelle nachträglich aus vorhandenen Datensätzen (zum Beispiel Museumsdaten) berechnet [36]. Die Methodik der Stichprobenerhebung kann einen signifikanten Einfluss auf die Modellgüte haben. Es wird zwischen regulärer, zufälliger, stratifizierter (equal stratified sampling) und proportional stratifizierter (proportional stratified sampling) Stichprobenerhebung unterschieden [26]. Studien zur Evaluierung der verschiedenen Methoden zeigen unterschiedliche Ergebnisse [55].

2.2.2 Statistische Modellformulierung

Die statistische Modellformulierung betrifft die Auswahl eines geeigneten Modellierungsansatzes, um Response und Umwelt zu verknüpfen. Die Eignung eines mathematischen Ansatzes hängt oft von der zu modellierenden Responsevariablen ab [21]. Moderne Verfahren stellen immer häufiger Ausnahmen dieser „Regel“ dar. Nichtparametrische Verfahren, wie zum Beispiel NPMR (Nonparametric Multiplicative Regression) [37], stellen grundsätzlich keine Bedingung an die Form der Responsekurve, aber auch parametrische Verfahren, wie zum Beispiel MaxEnt, setzen keine bestimmte Form voraus, sondern bestimmen diese aus den Daten.

Allgemeines mathematisches Modell

Aufgabe eines mathematischen Modells ist es, die vorhandenen Daten und das bisherige Wissen über die Verteilung einer Art mit den ausgewählten Umweltfaktoren in Beziehung zu stellen. Je nach Modellierungstechnik werden dazu die Parameter einer Funktion angepasst, Entscheidungsbäume erzeugt, oder eine Menge von Intervallen, die das Vorkommen der Art im Raum der Umweltvariaben beschreiben

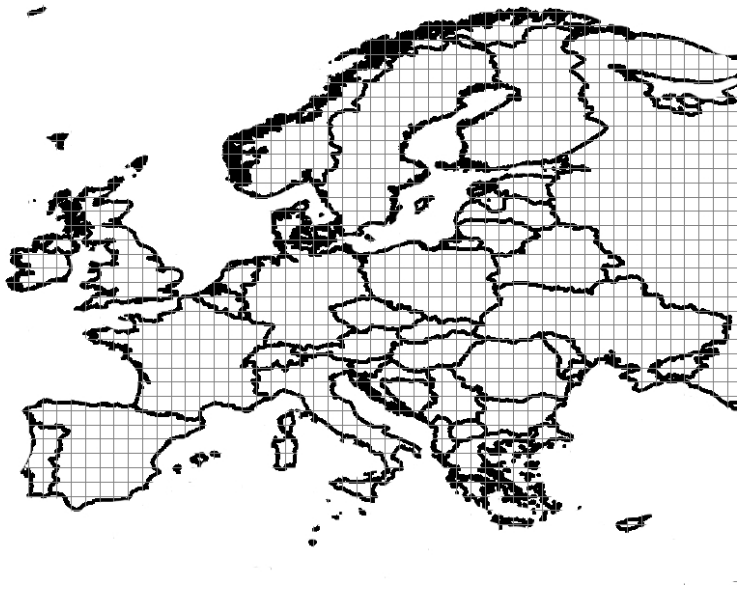


Abbildung 2.2: Beispielhafte Rasterisierung des geographischen Raumes mit einem 1°-Gitter (ca. 100km Rastergröße). Wasserflächen sind hier nicht mit Rasterzellen belegt. In den verfügbaren Rasterkarten werden Wasserzellen durch einen bestimmten Wert als NODATA-Zellen markiert und in der Berechnung der Modelle nicht betrachtet.

(Envelope-Verfahren), gelernt. Jeder dieser Ansätze kann formal als mathematische Funktion M aufgefasst werden. M wird während der Kalibrierungsphase durch ein Lernverfahren an die Daten angepasst.
Folgende Daten sind notwendig, um ein Habitatmodell zu kalibrieren:

- Stichproben der Vorkommen der Art (Präsenz- und eventuell Absenzpunkte) aus eine Untersuchungsgebiet

- Umweltvariablen und ihre Ausprägung an den Stichprobenpunkten

Das Untersuchungsgebiet dient als Grundlage für die Stichproben und ist die Basis für die spätere Vorhersage des Habitats (siehe 2.2.4).
Ein Punkt im Untersuchungsgebiet kann durch die Angabe von Längen- und Breitengrad identifiziert werden. Zur Vereinfachung werden Habitatmodelle meist nicht stetig auf dem Raum, sondern auf diskretisierten **Rasterkarten** berechnet. Dazu werden alle Koordinaten, die in eine Rasterzelle fallen, durch den Index oder die Koordinate des Zentrums dieser Rasterzelle identifiziert. Dieser diskretisierte Raum lässt sich als Menge X von Rasterzellen x_i auffassen.

$$X = \{x_i \mid i = 1, 2, \ldots, L\} \tag{2.1}$$

Der **Umweltzustand** wird als m-dimensionaler reellwertiger Funktionsvektor $\mathbf{f} : X \mapsto \mathbb{R}^m$ dargestellt. Die Komponenten $f_j(x), \forall j = 1, 2, \ldots, m$ repräsentieren die **Umweltvariablen**. Die **Modellfunktion**

$$M(\mathbf{f}(x)) : \mathbb{R}^m \mapsto \mathbb{R} \tag{2.2}$$

ordnet einem Umweltvektor $\mathbf{f}(x)$ eine **Habitatgüte** $g = M(\mathbf{f}(x))$ zu. Der Funktionswert g ist die modellierte Response einer Art oder Gesellschaft. Zur Berechnung von M werden Lernbeispiele benötigt. Die **Stichprobe** $sample \subseteq X$ wird zufällig aus dem Untersuchungsgebiet gezogen. Jedem Element $x \in sample$ wird eine Beobachtung $y : X \mapsto \{0, 1\}$ zugeordnet. Dabei ist $y(x) = 1$ für Rasterzellen in denen die Art oder Gesellschaft beobachtet wurde. Die Menge $ps = \{x \in sample \,|\, y(x) = 1\}$ wird **Präsenzstichprobe** genannt. Die Menge $as = \{x \in sample \,|\, y(x) = 0\}$ stellt die **Absenzstichprobe** dar. Eine Modellierungsmethode ist ein **Lernverfahren** $\mathcal{K}$, welches die Stichproben und Umweltvariablen als Eingabe nimmt und ein daran angepasstes Modell $M(f(x)) = \mathcal{K}(sample, \mathbf{f})$ ausgibt.

2.2.3 Kalibrierung

In der Kalibrierungsphase wird ein Modell an die Stichprobendaten angepasst, indem das, für den konkreten Modellansatz spezielle, Lernverfahren $\mathcal{K}$ angewandt wird. Ist eine Absenzstichprobe vorhanden, wird $\mathcal{K}$ als **Präsenz-Absenz-Verfahren** bezeichnet, sonst als **Präsenz-Verfahren**.
Die Kalibrierung von Modellen mit den in dieser Arbeit verwendeten Verfahren, Maximum-Entropie und Genetisches Programmieren, wird in den Kapiteln 2.3 und 2.4 beschrieben. Beide stellen Präsenzverfahren dar und benötigen keine Absenzstichprobe zur Modellerstellung, aber eventuell zur Modellevaluierung.

2.2.4 Vorhersage

Nach der Kalibrierung kann das Modell auf das Untersuchungsgebiet angewendet werden, indem die Modellfunktion auf allen Punkten der Rasterkarte ausgewertet wird. Jedem Punkt wird dadurch eine Vorkommenswahrscheinlichkeit zugewiesen. Dazu muss für jeden Punkt der Karte die Umwelt $\mathbf{f}(x)$ definiert sein.

2.2.5 Modellevaluation

Nach der Kalibrierungsphase sollte das Modell in Bezug auf folgende Eigenschaften bewertet werden:

- Anpassungsgüte (engl.: goodness-of-fit),
- Verfeinerung (engl.: refinement),
- Diskriminanz (engl.: discrimination).

Bei großen Stichproben, die auf zwei Datensätze aufgeteilt werden können (split sample [21]), oder bei Verfügbarkeit von zwei unabhängigen Stichproben, kann ein Teil der Daten zur Kalibrierung und der Rest zur Evaluierung herangezogen werden. Verfahren wie Kreuzvalidierung (KV), Jack-Knife (JK) oder Boostrapping werden bei kleinen Stichprobengrößen angewandt.
Ein einzelnes Gütemaß kann die Qualität eines Modell nicht ausreichend beschreiben, da kein Maß alle Eigenschaften eines Modells gleichzeitig erfassen kann. Daher werden in der Regel mehrere Maße zur Evaluierung herangezogen.

Anpassungsgüte

Die Anpassung des Modells wird in der Regel mit demselben Maß bewertet, das während der Kalibrierungsphase verwendet wurde [21], zum Beispiel R^2. R^2 (Definition 2) erfasst sowohl die Kalibrierung als auch die Verfeinerung des Modells und basiert auf der Likelihood $\mathcal{L}$ der Stichprobe. Wie gut das Modell an die Stichprobe angepasst ist, wird ermittelt, indem die Log-Likelihood der Stichprobe unter dem Modell berechnet wird und mit der Log-Likelihood eines Nullmodells verglichen wird. Die Berechnung der Likelihood ergibt sich aus Definition 1 angegeben.

Definition 1 Likelihood

Sei p_i die Wahrscheinlichkeit des i-ten Stichprobenpunktes unter einem Modell und y_i das Label dieses Punktes. Die Likelihood berechnet sich dann nach:

$$\mathcal{L} = \prod_{i=1}^{N} (p_i)^{y_i} \cdot (1 - p_i)^{(1-y_i)}, \qquad y_i \in \{0, 1\}, p_i \in [0, 1] \tag{2.3}$$

$$\mathcal{LL} = \ln(\mathcal{L}) = \sum_{i=1}^{N} y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i), \qquad y_i \in \{0, 1\}, p_i \in [0, 1] \tag{2.4}$$

Betrachtet man die Stichprobe als Ergebnis eines Zufallsprozesses, so stellt die berechnete Likelihood die Wahrscheinlichkeit dar, die Stichprobe zufällig genauso mit 0 und 1 zu labeln wie beobachtet.

Definition 2 R^2 nach Nagelkerke [39]

$\mathcal{LL}_{\mathcal{M}}$ bezeichne die logarithmierte Likelihood der kalibrierten Modells und $\mathcal{LL}_0$ die log-Likelihood des Nullmodells ($\forall i : p_i \equiv$ Prävalenz)

$$R^2 = \frac{1 - e^{-\frac{2}{N}(\mathcal{LL}_M - \mathcal{LL}_0)}}{1 - e^{\frac{2}{N}(\mathcal{LL}_0)}} \tag{2.5}$$

R^2 ist nach oben durch 1 beschränkt. In der Praxis ist der maximale Wert von R^2 kleiner als 1 [39, 11]. Durch die Normierung auf das maximale $R^2_{max} = 1 - e^{\frac{2}{N}(\mathcal{LL}_0)}$ wird eine Anpassung des Wertebereichs erreicht. R^2 kann auch negative Werte annehmen, wenn $\mathcal{LL}_M + N/2 < \mathcal{LL}_0$. Ein perfekt verfeinertes Modell weist allen Punkten nur den Wert 0 oder den Wert 1 zu. Werden alle Beobachtungen gleichzeitig auch korrekt vorhergesagt, so hat R^2 den Wert 1 [11]. Da R^2 konsistent zum Bestimmtheitsmaß der linearen Regression ist, kann es als Anteil der durch das Modell erklärten Variation interpretiert werden [39].

Diskriminanzmaße

Die Diskriminanz, die Unterscheidungsfähigkeit des Modells zwischen Präsenz und Absenz, kann durch verschiedene Gütemaße erfasst werden. Das Modell wird dazu als Klassifikator verwendet, um die Lern- bzw. Teststichprobe nach Vorkommen und Nichtvorkommen zu (re-)klassifizieren. Bei Modellen, die als Response jeden Wert aus dem Intervall [0,1] annehmen können, muss eine Entscheidung anhand eines Schwellenwertes getroffen werden. Da die Stichproben gelabelt sind, ist die richtige Klassifikation bekannt, und es kann gezählt werden, wie viele Stichprobenelemente richtig oder falsch klassifiziert wurden. Anhand dieser Information wird die Konfusionsmatrix aufgestellt, in welcher die Anzahl der richtig (a) und falsch positiv (c) bzw. richtig (b) und falsch negativ (d) klassifizierten Stichprobenelemente notiert wird. Für die Spaltensummen gilt $a + c = |ps|$ und $b + d = |as|$. Die Zeilensummen $a + b$ und $c + d$ entsprechen der Anzahl der positiv und negativ vorhergesagten Stichprobenelemente. Aus den Komponenten der Konfusionsmatrix lassen sich verschiedene Gütemaße berechnen, zum Beispiel Sensitivität, Spezifität, Cohen's Kappa, Max-Kappa und die Receiver-Operator-Charakteristik. In dieser Arbeit wird für

die Bewertung der Diskriminanz die Größe AUC, die Fläche unter ROC-Kurve, verwendet.
Der Wert von AUC lässt sich auf verschiedene Weisen berechnen. Nach der direkten Definition [23] kann AUC als Fläche unter der ROC-Kurve berechnet werden. Die ROC-Kurve lässt sich erzeugen, indem alle im Modell vorkommenden Wahrscheinlichkeiten als Schwelle verwendet werden und jeweils die Sensitivität $a/(a+c)$ und die Spezifität $d/(d+b)$ bestimmt werden [15]. Die Sensitivität wird dann als Funktion von (1-Spezifität) dargestellt und dadurch die ROC-Kurve generiert [22]. Die Fläche unter der so beschriebenen Kurve kann zum Beispiel durch numerische Integration berechnet werden.
AUC kann theoretisch Werte zwischen 0 und 1 annehmen. Ein Wert von 0,5 entspricht dabei einem zufälligen Modell. Werte kleiner als 0,5 können auftreten, wenn sich mindestens ein Punkt des Graphen unterhalb der Winkelhalbierenden befindet. Praktisch tritt dieser Fall nicht ein, da diese Punkte an der Winkelhalbierenden gespiegelt werden können, indem der Klassifikator (die 0-1 Entscheidung) umgekehrt interpretiert wird. Der AUC-Wert kann als Wahrscheinlichkeit interpretiert werden, dass bei einem zufälligen Paar von einem Präsenz- und einem Absenzpunkt dem Präsenzpunkt die höhere Wahrscheinlichkeit zugeordnet wird [15, 11].
Eine alternative Berechnung basiert auf der Eigenschaft, das AUC der skalierten Mann-Whitney-U-Statistik entspricht (Reineking in [11]). Dazu werden Präsenz und Absenzstichprobe in eine gemeinsame Reihenfolge gebracht und der mittlere Rang R_1 der Präsenzstichprobe bestimmt. Dieser geht in folgende Formel ein, in der N_1 der Anzahl der Elemente der Präsenzstichprobe und N der Gesamtanzahl von Stichprobenpunkten entspricht (Reineking in [11]):

$$AUC = \frac{1}{N - N_1}\left(R_1 - \frac{N_1 + 1}{2}\right). \tag{2.6}$$

In MAXENT wird ebenfalls die Verwandtschaft zur Mann-Withney-U-Statistik ausgenutzt. Die Sensitivität wird durch Auslassungsrate (Omission rate) ersetzt und (1-Spezifität) durch den Anteil der vorhergesagten Fläche (fractional predicted area, Anteil der Fläche, mit einer Wahrscheinlichkeit größer der Schwelle, an der Gesamtfläche)[48]. Zur Bewertung der vorgestellten Modelle wird diese Art der Berechnung verwendet, da die Werte in der Ausgabe von MAXENT direkt verfügbar sind. Alle alternativen Berechnungswege haben den Vorteil, dass kein Schwellenwert angewandt werden muss, sondern die vom Modell vorhergesagten, Wahrscheinlichkeiten verwendet werden.

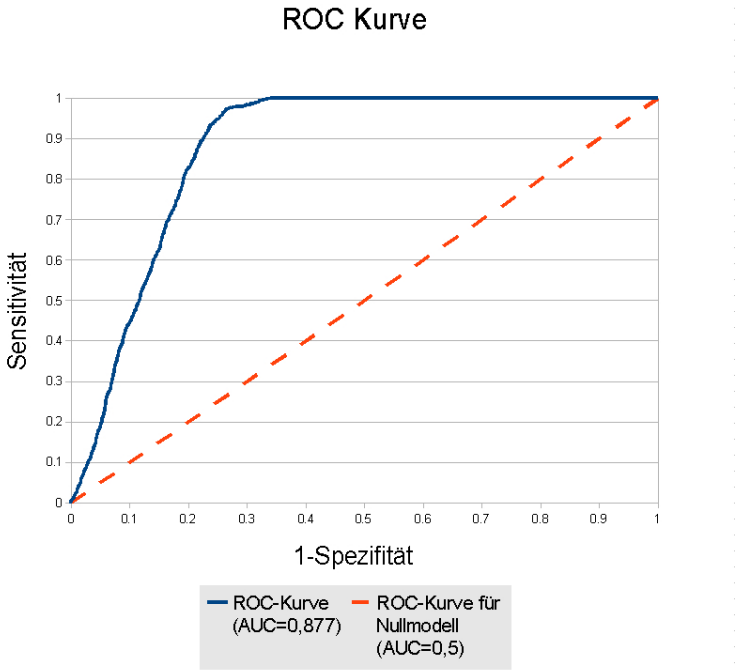


Abbildung 2.3: Beispiel für ROC-Kurven. Die Größe AUC wird als Fläche unter der Kurve berechnet.

Gütemaße im Präsenz-Szenario

Grundlegendes Problem, des hier betrachteten Präsenz-Szenarios, ist, dass die Diskriminierfähigkeit eines Modells nicht gemessen werden kann, da keine Absenzdaten zur Verfügung stehen. Um die Güte dieser Modelle trotzdem einzuschätzen, kann ausschließlich die Präsenzstichprobe verwendet werden. Eventuell kann eine Pseudo-Absenz-Stichprobe erhoben werden. Dies hat allerdings den Nachteil, dass nicht garantiert ist, dass alle Punkte in dieser Stichprobe echte Absenzpunkte sind. Durch falsche Absenzpunkte wird ein Fehler in die Berechnung des Gütmaßes eingebracht. Alle Gütemaße, die keine Diskriminierung der Stichprobe erfordern, können ohne große Einschränkungen im Präsenz-Szenario verwendet werden. Durch das Fehlen der Absenzstichprobe bzw. die dadurch unbekannte Prävalenz (Verhältnis von Präsenz- zu Absenzpunkten) ergeben sich allerdings Änderungen in der Berechnung. Die Berechnung der Likelihood vereinfacht sich, da der Term für die Absenz-Stichprobenpunkte immer 0 im Exponent besitzt, zu

$$\mathcal{L} = \prod_{i=1}^{N} p_i, \qquad y_i \in \{0,1\}, p_i \in [0,1] \tag{2.7}$$

$$\mathcal{LL} = \sum_{i=1}^{N} \ln(p_i), \qquad y_i \in \{0,1\}, p_i \in [0,1]. \tag{2.8}$$

Die Berechnung des Nullmodells für R^2 wird insofern verändert, dass allen Elementen aus *ps* der Wert 0,5 zuordnet wird, anstatt der Prävalenz. Dies hat zur Folge, dass die Normierung (Nenner von R^2) immer auf 0,75 erfolgt, wie durch die folgende Gleichungsfolge zu sehen ist.

$$\begin{aligned} 1 - e^{\frac{2}{N}\mathcal{LL}_0} &= 1 - e^{\frac{2}{N}\sum_{i=1}^{N}\ln(0,5)} &&= 1 - e^{\frac{2}{N}\ln(0,5)\sum_{i=1}^{N}1} \\ &= 1 - e^{\frac{2}{N}\ln(0,5)N} &&= 1 - e^{2\ln(0,5)} \\ &= 1 - e^{\ln(0,5^2)} &&= 1 - 0,5^2 = 0,75 \end{aligned} \tag{2.9}$$

Ein reines Präsenz-Gütemaß ist das Boyce'sche Verhältnis von vorhergesagter zu erwarteter Fläche (predicted-expected-ratio, PE-Ratio) (vgl. [27]). Hierbei wird gemessen, wie die Vorhersagegüte des Modells sich in bestimmten Bereichen der Wahrscheinlichkeiten verhält. Dazu werden b disjunkte Intervalle über dem Intervall [0,1] definiert. Für jedes Intervall i wird das Verhältnis von vorhergesagter zu erwarteter Fläche berechnet, indem der Anteil der Stichprobenelemente, die eine Wahrscheinlichkeit aus dem Intervall i zugeordnet bekommen, mit dem Anteil der Fläche, die dieses Intervall bedeckt, ins Verhältnis gesetzt wird (Definition 3).

Definition 3 PE-Ratio nach Boyce (vgl. [27])

Sei b die Anzahl der Intervalle, die auf [0,1] definiert sind. Sei $g_x = M(\mathbf{f}(x)$ die Response des Modells und es gelte $g \in [0,1]$. Für jedes Intervall $i \in \{1,2,\ldots,b\}$ berechnet sich die PE-Ratio aus

$$P_i = \frac{p_i}{\sum_{j=1}^{b} p_j}, \qquad p_i = |\{x \in ps \mid g_x \text{ liegt im Intervall } i\}| \tag{2.10}$$

$$E_i = \frac{a_i}{\sum_{j=1}^{b} a_j}, \qquad a_i = |\{x \in X \mid g_x \text{ liegt im Intervall } i\}| \tag{2.11}$$

$$F_i = \frac{P_i}{E_i} \tag{2.12}$$

Die Summen in den Nennern von P_i und E_i ergeben $|ps|$, die Stichprobengröße bzw. $|X|$ die Anzahl der Rasterzellen.

Da die PE-Ratio stark von der Intervallanzahl abhängt, haben Hirzel et al. [27] eine Erweiterung vorgeschlagen, die eine kontinuierliche Berechnung der PE-Ratio ermöglicht. Dazu werden die Intervalle nicht vorher festgelegt, sondern ein Gleitfensterverfahren angewandt. Ein Fenster der Breite w wird dabei iterativ über [0,1] bewegt und für die Fenstermitte jeweils die PE-Ratio berechnet. So entsteht eine glattere Kurve, welche einen guten Einblick in die Modellgüte erlaubt. Daraus kann die Perfomanz in verschiedenen Bereichen der Vorhersage bewertet werden.

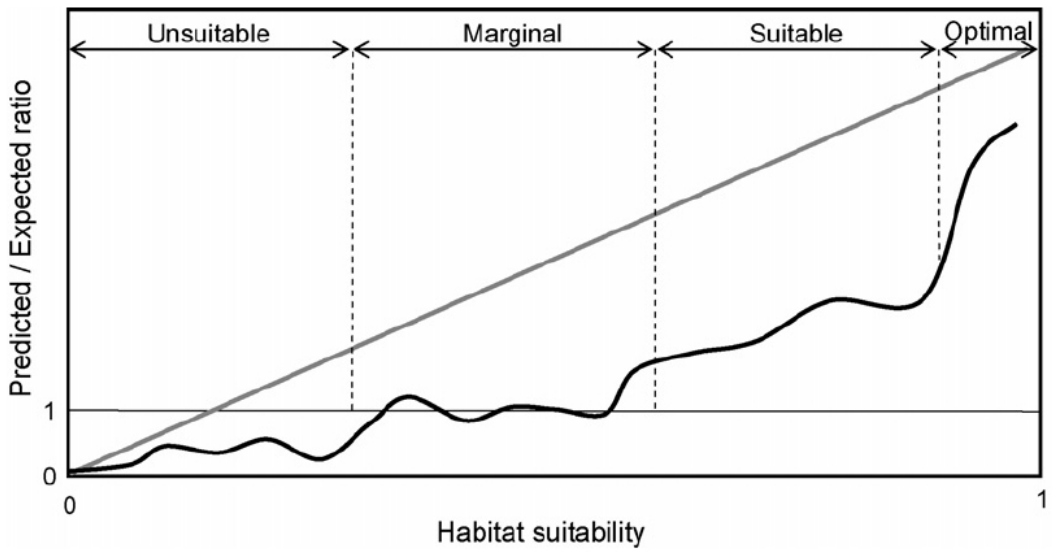


Abbildung 2.4: Beispiel für PE-Graphen, aus [27]. Auf der x-Achse wird die Habitatgüte, auf der y-Achse die PE-Ratio aufgetragen. Ein konstanter Funktionsverlauf von 1 entspricht einem komplett zufälligen Modell. Die Gerade stellt das optimale Modell dar. Ein gutes Modell zeigt einen ähnlichen Verlauf wie die schwarze Kurve.

Eine Gerade im PE-Graphen entspricht dem optimalen Modell, wobei die Steigung von den Daten abhängt. Ein komplett zufälliges Modell würde einen konstanten Wert von 1 erzeugen. Gute Modelle erzeugen eine ansteigende Kurve, während schlechte Modelle bei einem bestimmten Wert auf 0 abfallen. Der maximale Wert hängt vom Verhältnis $\frac{|X|}{|ps|}$ ab. Die Formel für die PE-Ratio lässt sich auflösen zu:

$$F_i = \frac{P_i}{E_i} = \frac{\frac{p_i}{\sum_{j=1}^{b} p_j}}{\frac{a_i}{\sum_{j=1}^{b} a_j}} = \frac{\frac{p_i}{|ps|}}{\frac{a_i}{|X|}} = \frac{p_i \cdot |X|}{a_i \cdot |ps|}. \tag{2.13}$$

Der absolute Wert wird durch das Verhältnis von p_i und a_i bestimmt. Die vorhergesagte Fläche p_i kann nie größer als die erwartete Fläche sein, da nie mehr Punkt aus der Stichprobe eine Wahrscheinlichkeit aus dem Intervall i bekommen können als insgesamt Punkte mit einer Wahrscheinlichkeit aus i belegt sind, so dass

$$p_i \leq a_i \Leftrightarrow \frac{p_i}{a_i} \leq 1 \tag{2.14}$$

gilt. Da das Verhältnis maximal den Wert 1 annehmen kann, ergibt sich das Maximum des gesamten Ausdrucks zu:

$$F_i \text{ ist maximal} \Leftrightarrow \frac{p_i}{a_i} = 1 \Rightarrow F_{i,max} = \frac{|X|}{|ps|}. \tag{2.15}$$

Für die Auswertung der in Kapitel 4 berechneten Modelle wurden R^2, AUC und das kontinuierliche Boyce-Maß verwendet.

2.3 Die Maximum-Entropie-Methode

Die Maximum-Entropie-Methode (MaxEnt) zur Berechnung von Habitatmodellen geht auf das Maximum-Entropie-Prinzip (ME-Prinzip) von Jaynes [30] zurück. Das ME-Prinzip kommt bei der Schätzung unbekannter oder nur teilweise bekannter Verteilungen zum Einsatz. Nach Jaynes soll die Schätzung einer unbekannten Wahrscheinlichkeitsverteilung das bisherige Wissen über diese Verteilung so gut wie möglich wiedergeben und daher die größtmögliche Entropie aufweisen. Es wird somit eine möglichst wenig strukturierte Verteilung gesucht, die alle bekannten Informationen wiedergibt.
Phillips, Dudik und Schapire (2004) wandten das ME-Prinzip erstmals an, um die Verteilung einer Art zu schätzen und dadurch ein Habitatmodell zu berechnen. Die von den Autoren erstellte Software MAXENT wurde unter anderem zur Modellierung von südamerikanischen Wäldern [50, 9], Herbariumsvorkommen [36], aber auch zur Modellierung von Artenreichtum [17] und zur Bewertung der Ausbreitung einwandernder Arten [14] verwendet. Der Vergleich verschiedener Methoden hat gezeigt, dass MaxEnt in der Regel bessere Ergebnisse erzeugt als andere Techniken [43, 24, 12].

2.3.1 Maximum-Entropie-Habitatmodelle

Die Funktion der **ME-Verteilung** kann hergeleitet werden, indem das ME-Prinzip als Optimierungsproblem formuliert wird und dieses mittels der Lagrange-Multiplikatoren-Methode gelöst wird [30].
Gesucht wird die Approximation $\widetilde{\pi}$ der unbekannten Verteilung π. Für die Approximation werden die schon bekannten Informationen zu π verwendet. Diese werden durch **Features** f vermittelt, genauer durch die Differenz der Erwartung $E_{\widehat{\pi}}[f]$ von den Erwartungswerten $E_{\widetilde{\pi}}[f]$. Die Differenzen

$$F_i = E_{\widehat{\pi}}[f] - E_{\widetilde{\pi}}[f] \stackrel{!}{=} 0 \tag{2.16}$$

werden als **Nebenbedingungen** für das Optimierungsproblem verwendet. Die Entropie $H(\widetilde{\pi}) = -\sum \widetilde{\pi} \ln \widetilde{\pi}$ geht als zu optimierende Eigenschaft in die Zielfunktion ein. Löst man das Optimierungsproblem, dass durch die **Zielfunktion**

$$L = H(\widetilde{\pi}) + \sum_i \lambda_i \cdot F_i \tag{2.17}$$

formuliert wird, so erhält man die **Gibbs-Verteilung**

$$\widetilde{\pi} = q_\lambda = \frac{e^{\sum_{j=0}^{m} \lambda_j \cdot f_j(x)}}{Z_\lambda} \tag{2.18}$$

(vgl. [30, 47]) mit

$$Z_\lambda = \sum_{x \in X} e^{\sum_{j=0}^{m} \lambda_j \cdot f_j(x)}. \tag{2.19}$$

Durch einen geeigneten Algorithmus, wie Generalized Iterativ Scaling oder Gradientenmethoden, können die Werte der λ_j berechnet werden.
Wendet man diesen Ansatz zur Berechnung von Habitatmodellen an, so wird versucht die Verteilung einer Art im Raum X zu beschreiben. Diese Verteilung im Raum entspricht der unbekannten Verteilung $\pi(x)$. Das bisherige Wissen wird durch die Stichprobe $ps \subset X$ und die Features f_i vermittelt. Die Features können sowohl Umweltvariablen sein, aber auch von diesen abgeleitete Größen, wie zum Beispiel das Quadrat einer Variablen, oder das Produkt zwischen zwei Umweltvariablen. Die Mittelwerte $E_{\widehat{\pi}}[f]$ werden für die Nebenbedingungen aus der Stichprobe und der Umwelt unter Annahme der Gleichverteilung ($\widehat{\pi}$) berechnet

$$E_{\widehat{\pi}}[f] = \frac{1}{n} \sum_{x \in ps} f_i(x). \tag{2.20}$$

Zur Kalibrierung der Verteilung $\widetilde{\pi}$ verwenden Phillips und Dudik eine sequentielle Abwandlung der Generellen Iterativen Skalierung (GIS). Dabei werden die λ_j nacheinander nach folgender Regel aktualisiert [49]:

$$\lambda_j = \lambda_j + \alpha, \qquad \alpha = \ln\left(\frac{E_{\widehat{\pi}}[f_j](1 - E_{q_\lambda}[f_j])}{(1 - E_{\widehat{\pi}}[f_j])E_{q_\lambda}[f_j]}\right). \tag{2.21}$$

Die Erwartungswerte $E_{q_\lambda}[f]$ der Gibbsverteilung werden dazu in jedem Schritt neu mit den aktuellen Parametern λ durch

$$E_{q_\lambda}[f] = \sum_{x \in X} q_\lambda \cdot f_i(x). \tag{2.22}$$

berechnet.

2.3.2 Interpretation der Maxent-Verteilung

Die durch die Kalibrierung enstandene Wahrscheinlichkeitsverteilung über dem geographischen Raum (dem Untersuchungsgebiet) weist in der Regel sehr kleine absolute Werte p_i (raw probabilites) auf. Diese Rohwahrscheinlichkeiten lassen sich als Anteil der modellierten Vegetation auf der Karte interpretieren. Dies ist ein Unterschied zu anderen Ansätzen, welche als Ergebnis eine Karte erzeugen, bei der jedem Punkt die Wahrscheinlichkeit zugeordnet wird, die Art dort anzutreffen. Diese Wahrscheinlichkeit ist der Parameter einer diskreten Zweipunktverteilung bezüglich Vorkommen-Nichtvorkommen.

Diese Interpretation als Zweipunktverteilungskarte wird in MaxEnt durch eine Transformation der räumlichen Verteilung erreicht, da eine direkte Berechnung nicht möglich ist [48]. In der Software MAXENT sind zwei verschiedene Transformationen zur Berechnung der Zweipunktwahrscheinlichkeiten implementiert: die kumulative Transformation und die logistische Transformation.
Die kumulative Transformation ersetzt jede Wahrscheinlichkeit durch die Summe aller kleineren Wahrscheinlichkeiten. Jedem Rasterpunkt x_i aus X wird durch

$$q_{\lambda,kumuliert}(x_i) = \sum_{x \in X | q_\lambda(x) \leq q_\lambda(x_i)} q_\lambda(x) \tag{2.23}$$

eine Wahrscheinlichkeit für Präsenz zugeordnet. Dabei tragen Rasterzellen aus ganz verschiedenen Regionen der Karte zu der Summe bei. Die Abhängigkeit von Habitatgüte und den Umweltvariablen geht dabei zum Teil verloren.
Die logistische Transformation erzeugt aus der räumlichen Verteilung für jeden Punkt x aus X eine Habitatgüte, die sich als Parameter der Zweipunktverteilung bezüglich Vorkommen-Nichtvorkommen interpretieren lässt. Die funktionale Form dieser Transformation ergibt sich aus der ME-Schätzung der gemeinsamen Wahrscheinlichkeitsverteilung $P(x, y = 1)$ [48], d.h. der Wahrscheinlichkeit, dass der Punkt x auftritt und ein Präsenzpunkt ist. Die Schätzung der gemeinsamen Wahrscheinlichkeit ist notwendig, da die Prävalenz $P(y = 1)$ nicht bekannt ist, aber durch die gemeinsame Verteilung $P(x, y = 1) = P(x|y = 1)P(y = 1)$ mitgeschätzt wird. Die funktionale Form der Transformation ist

$$q_{\lambda,logistisch}(x) = \frac{c \cdot q_\lambda(x)}{1 + c \cdot q_\lambda(x)}. \tag{2.24}$$

Die Konstante c wird aus der Entropie der geschätzen räumlichen Wahrscheinlichkeitsverteilung H_q mit $c = e^{H_q}$ berechnet. Diese Transformation wird von Phillips und Dudik nur genannt, aber bis auf eine kurze Erklärung nicht weiter hergeleitet. Eine wichtige Eigenschaft dieser Transformation ist, dass Punkte mit typischen Umweltkombinationen eine Wahrscheinlichkeit um 0,5 zugeordnet bekommen [48]. Andere Verfahren weisen diesen Punkten höhere Wahrscheinlichkeiten zu. Dies hat Auswirkungen auf die Modellbewertung durch die gebräuchlichen Gütemaße. R^2 vergleicht die Likelihood des Modells mit der Likelihood eines Nullmodells. Dieses Nullmodell belegt alle Stichprobenpunkte im Präsenz-Szenario mit 0,5. Die Bewertung eines MaxEnt-Modells, das 0,5 häufiger zuweist als andere Verfahren, führt zu einer kleineren Differenz ($\mathcal{LL} - \mathcal{LL}_0$) und dadurch zu einem kleineren R^2.

2.4 Genetisches Programmieren zur Habitatmodellierung

Genetisches Programmieren (GP) ist ein probabalistisches Optimierungsverfahren aus der Familie der evolutionären Algorithmen. Auf Grundlage der Prinzipien der Evolution, Reproduktion, Variation und Selektion, wird die Lösung eines Optimierungsproblems angenähert. Zur Familie der evolutionären Algorithmen gehören Genetische Algorithmen (GA), Evolutionsstrategien (ES), Evolutionäres Programmieren (EP) und Genetisches Programmieren (GP). In der Habitatmodellierung wurde GP schon eingesetzt. Das System GARP (Genetic Algorithm for Ruleset Prediction, [54, 4]) lernt zum Beispiel eine Menge von Regeln, die die Verteilung einer Art beschreiben. McKay [38] und Shan [53] verfolgen ähnliche Ansätze. GARP hat sich stärker etabliert und wurde mit anderen Modellierungstechniken verglichen. In den entsprechenden Studien [47, 46, 12] zeigt GARP gute, aber nicht optimale Ergebnisse.
In Kapitel 2.2.2 wurde dargestellt, dass sich Habitatmodelle durch eine Funktion beschreiben lassen, die einem beliebigen Umweltvektor eine Habitatgüte zuweisen. In der Regel stammt diese Funktion aus vorherigen theoretischen Überlegungen. Genetisches Programmieren lernt beliebige Funktionen und kann die funktionalen Zusammenhänge zwischen Umwelt und Responsevariable erfassen. Dabei ist GP nicht an eine Klasse von Funktionen gebunden, sondern hält eine Menge möglicher Lösungen bereit und wählt aus diesen gute Lösungen aus. Die selektierten Lösungen werden zufällig verändert, um eine Verbesserung zu erreichen.
Um mit MaxEnt vergleichbar zu bleiben arbeitet das vorgestellte genetische Programm (HabitatGP) ohne Absenzdaten. Die Güte verschiedener Individuen wird analog zu MaxEnt anhand der modellierten räumlichen Verteilung ermittelt und durch die Entropie und die Erwartungswerte der Verteilung erfasst.
In diesem Abschnitt wird der Aufbau des HabitatGP beschrieben. Die Ergebnisse der Testläufe werden in Kapitel 5 vorgestellt.

2.4.1 Algorithmus

Das HabitatGP implementiert folgenden Algorithmus:

1. Initialisiere die Population zufällig.

2. Bewerte die Fitness der Initialpopulation
3. Wähle zufällig nach einer bestimmten Methode Individuen aus, die fitter sind als andere.
4. Variiere diese Individuen durch
   a) Reproduktion
   b) Rekombination
   c) Mutation.
5. Bewerte die Fitness der neuen Individuen.
6. Ist das Abbruchkriterium nicht erfüllt, gehe zu 3. .
7. Gebe das beste gefundene Individuum als bisher beste Lösung des Optimierungsproblems aus.

Dieser Algorithmus ist an den allgemeinen Algorithmus, wie er für evolutionäre Lösungsverfahren häufig angewandt wird (vgl. [8]), angelehnt. Ein Flussdiagramm zum genauen Ablauf des HabitatGPs befindet sich im Anhang (Abbildung III.1).

2.4.2 Repräsentation der Individuen

Jedes Individuum stellt eine mögliche Modellfunktion M dar. Diese wird durch eine lineare Abfolge von Registeroperationen repräsentiert. Diese Darstellung des Genoms (Genotyp) wird unter dem Begriff „lineares genetisches Programmieren“(LGP) zusammengefasst. Brameier [8] gibt eine ausführliche Beschreibung von LGP.
Eine Registeroperation besteht aus einer linken und rechten Seite. Die linke Seite enthält nur das Zielregister. Die rechte Seite besteht aus einer mathematischen Operation und den Operanden-Registern. Das Zielregister darf auf der rechten Seite auftreten. Die Form der rechten Seite hängt von der verwendeten Funktionsmenge ab und kann binäre Operationen, unäre Operationen oder auch Bedingungen (IF-THEN-ELSE) enthalten. Das Ergebnis der Berechnung der rechten Seite wird in das Zielregister geschrieben, welches dabei überschrieben wird. Die Auswertung der Individuen erfolg linear von oben nach unten, indem alle Anweisungen nacheinander ausgeführt werden. Eine genauere Beschreibung zur Auswertung der Individuen wird im Abschnitt zur Genotyp-Phänotyp-Abbildung (siehe Kapitel 2.4.5) gegeben. Die Komplexität der Individuen wird durch die Anzahl der Operationen und verfügbaren Register bestimmt. Nähere Ausführung dazu finden sich in Kapitel 2.4.4. Ein Beispielindividuum ist in Abbildung 2.5 gegeben.

```
...
r[2] = r[5] + r[9]
r[1] = r[12] / r[2]
r[1] = Exp(r[1])
r[1] = Power(r[2])
...
```

Abbildung 2.5: Beispiel: Ausschnitt eines Individuengenoms

2.4.3 Funktionsmenge

In der Funktionsmenge werden die Operationen definiert, die zum Aufbau einer Anweisung des LGP verwendet werden können. Sie bilden zusammen mit der Menge der Konstanten-Register und den Input-Registern das Alphabet, aus dem das Genom der Individuen aufgebaut wird. Die Auswahl der Basisfunktionen ist ein kritischer Punkt in der Konstruktion des GP, da durch sie der Suchraum und damit die Menge und Art der möglichen Lösungen definiert wird. So können Lösungen, die nur auf die Addition als Funktion zurückgreifen, nur lineare Zusammenhänge zwischen Umwelt und Responsevariable lernen und sind in ihrer Mächtigkeit stark beschränkt. In der ökologischen Theorie wird angenommen, dass Responsekurven idealerweise einer Normalverteilung entsprechen, also symmetrische Glockenkurven sind. In der Natur werden Abweichungen von dieser Form beobachtet. Dort finden sich auch schiefe unimodale oder bimodale Responsekurven (zum Beispiel unter Konkurrenz). Die Menge von Basisfunktionen sollte die Möglichkeit bieten, die meisten Typen von Responsekurven zu erfassen.
Unverzichtbare Elemente der Funktionsmenge sind Addition und Subtraktion, da durch sie die einfachen linearen Zusammenhänge erfasst werden.
Die Multiplikation ermöglicht es, Lösungen zu finden, in denen Variablen einfach gewichtet sind bzw. in denen Variablen in höhere Potenzen durch mehrfache Multiplikation eingehen können (Exponentenbildung). Durch die Multiplikation von zwei Variablen können Interaktionen modelliert werden. Um unimodale, normalverteilungsähnliche Lösungen zu bekommen wird zusätzlich die Exponentialfunktion benötigt.
Durch Kombination der verschiedenen Basisfunktionen können komplexere Funktionen gebildet werden, die ein großes Spektrum von möglichen Zusammenhängen zwischen Umwelt und Responsevariable erfassen. Alle verwendeten Funktionen sind in Tabelle 2.1 zusammengefasst. Nicht mit einbezogen wurden der Logarithmus, die Wurzel und die trigonometrischen Funktionen. Im GP ist es ebenfalls denkbar bedingte Verzweigungen (IF-THEN-ELSE) mit einzubeziehen. Eine bedingte Verzweigung entspräche einer Unstetigkeit innerhalb der Funktion. Da das HabitatGP

geschlossene Ausdrücke berechnen soll, sind bedingte Verzweigungen ebenfalls nicht implementiert.
Um numerische Probleme soweit wie möglich zu vermeiden, sind alle Operationen numerisch abgesichert. Sollte das Ergebnis einer Operation größer als die maximal darstellbare Fließkomma-Zahl des Systems d_{max} sein, so wird das Ergebnis auf diese zurückgesetzt. Funktionen, die schnell große Zahlen erzeugen, wie e^x und x^y, sind auf Exponenten kleiner 10 beschränkt.

Funktion	numerischer Schutz
$r = x + y$	Falls $x + y > d_{max}$, dann $r = d_{max}$
$r = x - y$	Falls $x - y < d_{min}$, dann $r = d_{min}$
$r = x \cdot y$	Falls $x \cdot y > d_{max}$, dann $r = d_{max}$
$r = x/y$	Falls $x/y > d_{max}$ oder $y = 0$, dann $r = d_{max}$
$r = x^y$	Falls $y > 10$, dann $r = d_{max}$
$r = x^2$	
$r = e^y$	Falls $y > 10$, dann $r = d_{max}$

Tabelle 2.1: Verwendete Basisfunktionen und deren numerischer Schutz, um Überläufe so weit wie möglich zu verhindern.

2.4.4 Charakterisierung des Suchraums

Der Suchraum S wird durch die möglichen Genotypen der Individuen aufgespannt. Jedes Individuum besteht aus einer bestimmten Menge von Registeroperationen, die in einer bestimmten Reihenfolge auftreten. Jede dieser Registeroperationen schreibt in ein Zielregister und liest aus zwei beliebigen Registern die Operanden aus. Außerdem kann eine beliebige Operation aus der Menge der Basisfunktionen verwendet werden. Mathematisch lässt sich der Suchraum somit als Kreuzprodukt aus den Mengen beschreiben, die jeweils alle möglichen Ausprägungen einer dieser Komponenten beinhalten. Eine Übersicht ist in Tabelle 2.2 gegeben.

Abkürzung	Beschreibung
R	Die Menge aller Register
RR	Die Menge der Rechenregister.
KR	Die Menge der Konstantenregister.
IR	Die Menge der Inputregister.
OP	Die Menge der möglichen Operationen

Tabelle 2.2: Abkürzungen der Komponenten des Suchraums

Eine Registeranweisung a sei ein Element der Menge aller möglichen Registeranweisungen:

$$a \in A = RR \times R \times R \times OP, \tag{2.25}$$

mit $R = RR \cup KR \cup IR$.
Ein Individuum ind ist eine Folge von l beliebigen Registeranweisungen: $ind \in A^l$.
Der Suchraum S sei die Menge aller möglicher Individuen. Ist die Länge l der Individuen nicht nach oben beschränkt, lässt sich die Größe des Suchraums nicht berechnen. Sei l durch l_{max} beschränkt, so stellt sich der Suchraum als

$$S = \bigcup_{k=0}^{l_{max}} A^k \tag{2.26}$$

dar.
Die Mächtigkeit $|S|$ des Suchraums berechnet sich aus:

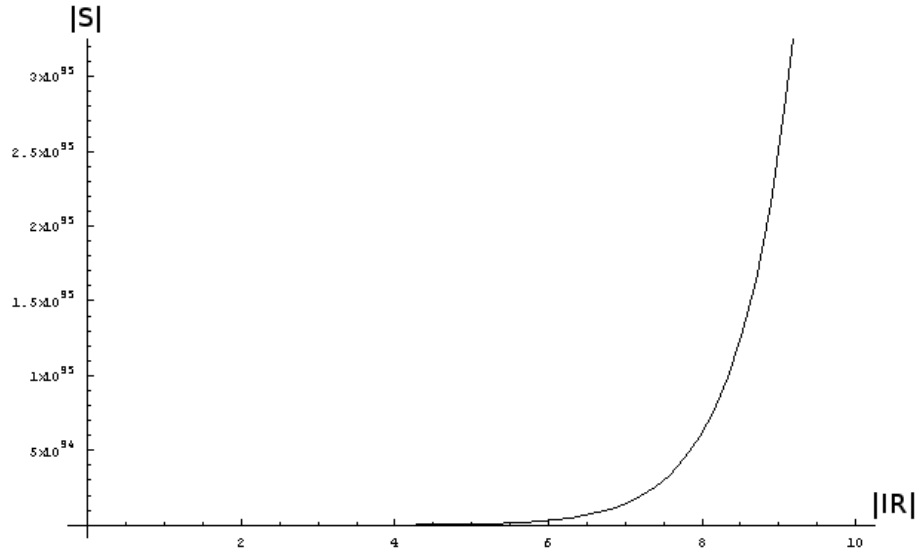
$$|S| = \sum_{k=0}^{l_{max}} |A|^k \tag{2.27}$$

$$= \sum_{k=0}^{l_{max}} (|RR| \cdot |R| \cdot |R| \cdot |OP|)^k \tag{2.28}$$

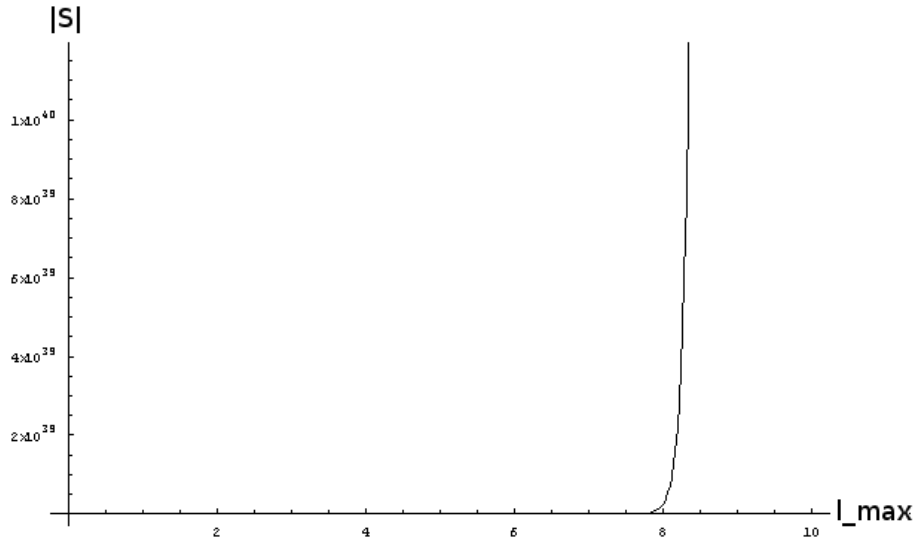
$$= \sum_{k=0}^{l_{max}} (|RR| \cdot |R|^2 \cdot |OP|)^k \tag{2.29}$$

$$= \sum_{k=0}^{l_{max}} (|RR| \cdot (|RR| + |KR| + |IR|)^2 \cdot |OP|)^k \tag{2.30}$$

Auf dieser Grundlage kann $|S|$ abgeschätzt werden. In Abbildung 2.6 ist für eine konkrete Parameterkombination die Abhängigkeit des Suchraums von der Anzahl der Variablen und der maximalen Länge angegeben.Im HabitatGP ist keine Längenbeschränkung implementiert, da so die Wahrscheinlichkeit steigt, gute Individuen zu finden, auch wenn sie länger sind.



(a) Wachstum des Suchraums in Abhängigkeit von der Variablenanzahl mit $l_{max} = 20, |RR| = 10, |KR| = 10, |OP| = 7$



(b) Wachstum des Suchraums in Abhängigkeit von der Individuenlänge mit $|IR| = 10, |RR| = 10, |KR| = 10, |OP| = 7$

Abbildung 2.6: Wachstum des Suchraums in Abhängigkeit der Variablenanzahl (a) und der Individuenlänge (b).

2.4.5 Genotyp-Phänotyp-Abbildung

Der Genotyp der Individuen beschreibt die funktionale Form des gelernten Modells. Um dieses Bewerten zu können, muss das Modell auf das Untersuchungsgebiet angewandt und so die geographische Wahrscheinlichkeitsverteilung berechnet werden. Die geographische Verteilung der modellierten Art ist der Phänotyp.

Definition 4 Genotyp-Phänotyp-Abbildung

Sei $x_i \in X$ und G_{ind} das Genom des zu bewertenden Individuums *ind*. Die Funktion `eval` wertet das Genom mit der Umwelt $\mathbf{f}(x_i)$ als Eingabe aus. Durch `eval` wird die Modellfunktion simuliert und die Vorkommenswahrscheinlichkeit p_i an Punkt x_i berechnet.

$$p_i = M_{ind}(\mathbf{f}(x_i)) = eval(G_{ind}, \mathbf{f}(x_i)) \tag{2.31}$$

Das Genom jedes Individuums enthält Introns, das sind Codeabschnitte, deren Auswertung das Ergebnis nicht verändert. Diese Introns können entfernt werden, bevor das Genom ausgewertet wird. Brameier schlägt einen einfachen Algorithmus vor, der in O(n) alle Introns findet [8]. Dieser ist im HabitatGP implementiert und der Auswertung vorgeschaltet (siehe Abschnitt 2.4.9).
Für die Genotyp-Phänotyp-Abbildung wird für jeden Rasterpunkt zuerst der Vektor der dort herrschenden Umwelt erstellt. Dieser wird zusammen mit dem (prozessierten) Genom des aktuellen Individuums `ind` an die Funktion `double_v eval(individuum_v ind, double_v envir, bool verbose )` übergeben. Ein Array vom Typ double enthält die Register. `eval()` belegt Inputregister mit der Umwelt, initialisiert die Rechenregister und die Konstanten. In einer Schleife wird jede Anweisung des Genoms ausgewertet und die Registerbelegung entsprechend geändert. Nachdem alle Registeroperationen ausgewertet wurden, wird die gesamte Registerbelegung zurückgegeben. Die aufrufende Funktion kann das Ergebnis aus einem festgelegtem Ausgaberegister ablesen. Nachdem alle Rasterpunkte ausgewertet wurden, werden die Ergebnisse auf die Summe über alle Punkte normiert, so dass eine Wahrscheinlichkeitsverteilung entsteht.

2.4.6 Zielfunktion

Mit der Zielfunktion wird mathematisch formuliert, welche Eigenschaften das optimale Modell haben soll. Zum Beispiel sollte die Stichprobe gut vorhergesagt werden und die Verteilung gleichzeitig eine große Entropie besitzen. Diese multiplen Kriterien müssen für die Fitnessberechnung zusammengebracht werden. Dies kann zum Beispiel durch eine aggregierte skalare Fitnessfunktion erreicht werden. Dabei werden alle Ziele zuerst einzeln bewertet und dann in einer Linearkombination zu einem Fitnesswert verrechnet. Die optimale Lösung soll die Fitness Null besitzen. Das hat zur Folge, dass die einzelnen Kriterien ebenfalls so formuliert werden müssen, dass die optimale Lösung in diesem Kriterium gegen Null strebt. Ich möchte mich hier auf die Verwendung einer aggregierten Fitnessfunktion beschränken, da diese die einfachste Möglichkeit darstellte die Fitness von Individuen zu bestimmen.
Die Fitnessbewertung basiert auf einem Hitkriterium und Eigenschaften der räumlichen Wahrscheinlichkeitsverteilung. Zusätzlich wird ein Term in die Fitnessfunktion eingebracht, der die Anzahl der in der Funktion verwendeten Prädiktorvariablen bewertet.

Berechnung von Hits

Die einfachste Möglichkeit, die Anpassung des Modells an die Daten zu messen, ist ein Hit-Kriterium. Dazu wird die zu bewertende Lösung auf die Stichprobenpunkte angewandt und bewertet, wie vielen eine hohe Wahrscheinlichkeit zugeordnet wird. Ein solches Hit-Kriterium kann implementiert werden, indem die Wahrscheinlichkeit jedes Stichprobenpunkts berechnet wird und mit einem Schwellenwert Präsenz und Absenz bestimmt wird. Da die Verteilung über dem geographischen Raum normiert ist, können verschiedene Modell unterschiedliche absolute Werte für die Stichprobenpunkte besitzen, so dass ein einheitlicher Schwellenwert in dieser Form nicht zu realisieren ist. Transformiert man die Wahrscheinlichkeiten vorher linear auf das Intervall [0,1] lässt sich ein einheitlicher Schwellenwert für alle Modelle festlegen.

Definition 5 Hit-Kriterium

Sei p_i^t die linear tranformierte Wahrscheinlichkeit des Stichprobenelements x_i und $th \in [0,1]$ eine Schwelle. $\mathbf{1}_{p_i^t > th}$ nimmt den Wert 1 an, wenn die Bedingung $p_i^t > th$ erfüllt ist, ansonsten 0.

$$Hits = \sum_{x_i \in ps} \mathbf{1}_{p_i^t > th} \tag{2.32}$$

Eigenschaften der Verteilung

Da mit dem HabitatGP eine Wahrscheinlichkeitsverteilung berechnet werden soll, können auch die Eigenschaften der Verteilung genutzt und zur Bewertung der Individuen verwendet werden. Ähnlich zu Maximum-Entropie können Erwartungswerte und Varianzen aus der Modellverteilung berechnet und mit den empirischen Werten (aus der Stichprobe) verglichen werden. Im Gegensatz zu Maxent ist aber nicht genau bekannt, wie die Variablen als Features in das Modell eingehen, so dass diese Modelleigenschaften nicht für jedes Feature, sondern nur für jede Variable berechnet und verglichen werden. Das optimale Modell entspricht dann im Erwartungswert und der Streuung der Erwartung der empirischen Verteilung, wie sie durch die Stichprobe beschrieben wird. Auf diese Weise werden Informationen zur Lage und Breite der Verteilung für den Lernprozess nutzbar gemacht.
Neben den Erwartungswerten und Varianzen bietet sich an die Entropie der Verteilung mit in die Betrachtung einzubeziehen. Diese soll maximiert werden. Die daraus resultierende Verteilung sollte somit die Eigenschaften der empirischen Verteilung wiederspiegeln und unbekannte Umweltkombinationen möglichst einfach erklären (Generalisieren).
Für die Formulierung der Zielfunktion wird die Abweichung Δ vom Optimalwert berechnet.

Definition 6 Abweichung von den Erwartungswerten, Varianzen und der Entropie zur Bewertung der Fitness der Individuen

$$\Delta_E = \sum_{i}^{\#Variablen} \frac{\widehat{\pi}[f_i] - \widetilde{\pi}[f_i]}{\widehat{\pi}[f_i]} \tag{2.33}$$

$$\Delta_{Var} = \sum_{i}^{\#Variablen} \frac{\widehat{\pi}[f_i]^2 - \widetilde{\pi}[f_i]^2}{\widehat{\pi}[f_i]^2} \tag{2.34}$$

$$\Delta_H = \frac{H_{max} - H}{H_{max}}, \text{ mit } H_{max} = \ln(|X|) \tag{2.35}$$

2.4.7 Selektion

In jeder Generation werden $2 \cdot$ `n_tour` Turniere der Größe 2 durchgeführt. Vier Turnierteilnehmer werden zufällig aus der Population gezogen, zwei für das erste und zwei für das zweite Turnier. Die Gewinner der zwei Turniere werden zur Rekombination (siehe 2.4.8) herangezogen und zwei Nachkommen erzeugt. Die Fitness

der Nachkommen wird berechnet und die Verlierer der Turniere ersetzt. Dies wird `n_tour`-mal wiederholt. Dabei können auch in der aktuellen Generation erzeugte Individuen wieder in ein Turnier eingehen. Durch diese Verschachtelung wird die Fortschrittsrate erhöht.

2.4.8 Variation

Individuen können auf verschiedenen Ebenen variiert werden. Während durch Mutationen zufällige Änderungen innerhalb eines Individuum vorgenommen werden, können durch Rekombination Teile der Genome auch zwischen Individuen ausgetauscht werden. Sowohl Mutation als auch Rekombination werden auf dem nichtprozessierten Genom durchgeführt, da so günstige „Gene“, die in den Introns vorhanden sind könnten, potentiell verwendet werden können.

Mutation

Bei Mutationen kann man zwischen Mikro- und Makromutationen unterscheiden. Während erstere die Operation (z.B. $+ \rightarrow exp()$) einer Anweisung oder ein Register verändern (z.B. r[1] $\rightarrow$ r[34]), fügen letztere eine Anweisung an einer zufälligen Stelle ein oder löschen eine zufällige Anweisung.
Für die Durchführung einer Mikromutation wird zufällig eine Anweisung ausgewählt. Danach wird zufällig die zu mutierende Komponente ermittelt (Target-Register, linkes, rechtes Register oder die Operation) und durch eine letzte Zufallsauswahl der neue Zustand der mutierten Komponenten bestimmt.
Im HabitatGP werden bei jedem Aufruf der Methode `mutate() n_mut` Versuche unternommen mit der Wahrscheinlichkeit `mutwsk`, eine Mutation in das Genom einzuführen. Die Auswahl zwischen Insertion, Deletion und Mikromutation wird durch eine gleichverteilte Zufallszahl getroffen, d.h. $\frac{2}{3}$ aller Mutationen sind Makromutationen.

Rekombination

Die Rekombination ermöglicht einen Austausch von genetischer Information zwischen Individuen, in der Hoffnung, dabei durch die Kombination vorteilhafter „Gene“ ein besseres Individuum zu bekommen. Rekombination kann verschieden reali-

siert werden. Im HabitatGP wird ein 2-Punkt-Crossover verwendet. Dabei wird für jedes Individuum zufällig ein Bereich des Genoms ausgewählt. Dieser wird durch den ausgewählten Bereich des jeweils anderen Individuums ausgetauscht. Da in HabitatGP generell keine Längenbeschränkung der Individuuen besteht, wird keine weitere Kontrolle der resultierenden Individuen durchgeführt.

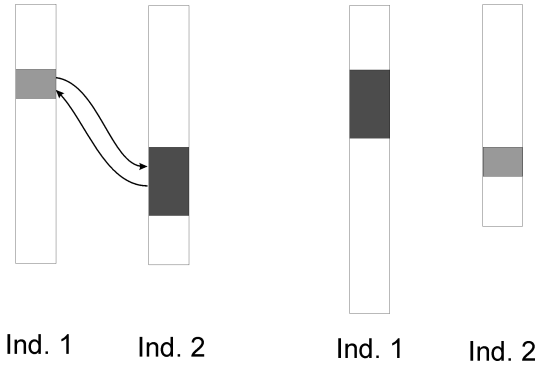


Abbildung 2.7: Schematische Darstellung des 2-Punkt-Crossover. Die beiden markierten Flächen werden für jedes Individuum zufällig ausgewählt. Diese beiden Teile der Genome werden ausgetauscht und dadurch zwei neue Individuen erzeugt.

2.4.9 Implementierungsdetails

Die Implementierung des HabitatGP erfolgte in C++.

Registerrepräsentation

Die Register, die im LGP verwendet werden, um Inputdaten und Zwischenergebnisse zu speichern, werden als einfaches `long double` Array implementiert. Die Registerbelegung beginnt immer mit den Inputregistern. Darauf folgen die Rechenregister. Das erste Rechenregister nach den Inputregistern ist das Ausgaberegister, das am Ende der Berechnung das Ergebnis enthält. Die Rechenregister werden für die Auswertung immer mit 0 initialisiert. Danach folgen die Konstantenregister. Während der Auswertung einer Funktion dürfen nur die Rechenregister als Targetregister verwendet werden.

Verwendete Datentypen

Die Individuen werden als `structs` gespeichert. Die Struktur `ind_s` fasst das Genom und zusätzliche Informationen zur Fitness des Individuum zusammen. `individuum_v` ist ein selbstdefinierter Datentyp, ein Vektor, in welchem die Registeranweisungen gespeichert werden (`typedef vector < instruction_s > individuum_v`). Die Struktur `instruction_s` speichert Registeroperationen.

```
struct ind_s{
        individuum_v genome;
        double fit;
        bool eval;
        long double summe;
        int hits;
        unsigned int usedFeatures;
        unsigned int usedOps;
        long double gesamt_diff;
        long double entropie;
        vector<long double> fitnesses;
};
```

Abbildung 2.8: Struktur der Individuen: fit enthält die Gesamtfitness. Die boolsche Variable `eval` markiert, ob das Individuum schon bewertet wurde. Die Variable `summe` speichert die Summe zur Normierung der, durch dieses Individuum entstandenen, Verteilung ab. Alle anderen Attribute werden zur Berechnung der Fitness herangezogen. Der Vektor `fitnesses` enthält alle Komponenten der Fitness gesammelt.

```
struct instruction_s{
        unsigned int target;
        unsigned int left;
        unsigned int right;
        unsigned int op;
};
```

Abbildung 2.9: Struktur der Registeroperationen

Auch unäre Operationen werden durch diese Struktur dargestellt. Bei der Auswertung von unären Operationen wird nur `left` als Operandenregister verwendet.

Die Datenstruktur für eine Population ist ein Vektor von Individuen und wird durch `typedef vector < ind_s > population_v;` definiert. Die Umweltdaten werden jeweils als separate Karte eingelesen und in Vektoren gespeichert. `maps_v maplist` enthält alle Umweltkarten. Eine Karte ist ein Vektor von Zeilen der Karte. Jede Zeile ist ein Vektor vom Typ `double`. Der Typ `double_v` wird an anderer Stelle im Programm wiederverwendet, zum Beispiel in der Methode `eval()` zur Verwaltung der Register.

```
typedef vector < long double > double_v; //Eine Reihe der Umweltkarte
typedef vector < double_v > map_v; //Vektor von Reihen = Eine Karte
typedef vector < map_v > maps_v; //Vektor von Karten
```

Abbildung 2.10: Struktur der Registeroperationen

Alle Karten haben gemeinsame Eigenschaften. Diese werden im Programm in der Variable `map_s map_attributes` verwaltet. Der Typ `map_s` speichert die Georeferenzierung (Koordinaten der Kartenecken, Größe der Rasterzellen) und die Anzahl der Spalten und Zeilen der Rasterkarte. `nodata_value` enthält den Zahlenwert, der undefinierte Rasterzellen markiert (zum Beispiel Wasserflächen).

```
struct map_s{
       unsigned int ncols;
       unsigned int nrows;
       double xllcenter;
       double yllcenter;
       double cellsize;
       double nodata_value;
};
```

Abbildung 2.11: Aufbau der Struktur für die Karteneigenschaften

Die Stichprobe ist eine Menge von Punkten der Karte repräsentiert durch einen Vektor von Präsenzpunkten `vector< pp_s > stichprobe`. Ein Präsenzpunkt wird durch folgende Struktur definiert:

```
struct pp_s{
       double x;
       double y;
       bool inMap;
       int pos_trunc;
};
```

Abbildung 2.12: Struktur der Samplepunkte

Da das HabitatGP nur aus Präsenzdaten lernt, wird eine Repräsentation von Absenzpunkte nicht benötigt. Die Variablen `x` und `y` speichern die räumlichen Koordinaten der Punkte (Längen und Breitengrad) und nicht den Index auf der Karte. Dieser wird später im Programmverlauf durch die folgende Transformation berechnet:

$$index(y) = nrows - \lfloor((y - yllcenter)/cellsize)\rfloor - 1 \tag{2.36}$$

$$index(x) = \lfloor(x - xllcenter)/cellsize\rfloor. \tag{2.37}$$

Strukturelle Introns

Um die Auswertungsdauer der Individuen zu verringern, werden strukturelle Introns vor der Auswertung zeitweise aus dem Genom entfernt. Der hier vorgestellte Algorithmus basiert auf einem Algorithmus, wie er von Brameier [8] beschrieben wurde. Strukturelle Introns sind Registeranweisungen, die keinen Einfluss auf das Ergebnis der Berechnung haben. Diese Anweisungen lassen sich erkennen, da sie ein Targetregister belegen, das im weiteren Verlauf der Berechnung nicht mehr verwendet wird. Der erste Schritt besteht darin, die letzte Anweisung zu finden, die das Ausgaberegister belegt. Das Genom wird dazu rückwärts durchsucht. Der gesamte Algorithmus enthält folgende Schritte:

1. Finde die letzte Anweisung a, die das Ausgaberegister als Target hat.
2. Lösche alle Registeranweisung, die nach a folgen.
3. Speichere die Operandenregister der aktuellen Registeranweisung in einer Menge `memory`.
4. Gehe eine Anweisung nach oben. Ist der Anfang des Genoms erreicht, breche den Algorithmus ab.
5. Ist das Targetregister in `memory` enthalten?

6. Wenn ja, dann lösche das Targetregister aus `memory` und gehe zu 3..
7. Wenn nein, lösche diese Anweisung und gehe zu 4. .

Das so prozessierte Genom ist signifikant kürzer als vorher und die Auswertung kann effektiver berechnet werden.

Fitnessberechnung eines Individuums

Die Fitnessberechnung eines Individuum setzt vorraus, dass die Wahrscheinlichkeitsverteilung, die durch dieses Individuum bestimmt wird, bekannt ist (Genotyp-Phänotyp-Abbildung).
Die Auswertung wird von der Methode `eval()` durchgeführt. `eval()` bekommt als Eingabe das effektive Genom des Individuums *ind* und einen Umweltvektor übergeben. Dieser Umweltvektor wird verwendet, um die Inputregister zu belegen. Die Rechenregister werden mit 0 initialisiert und die Konstantenregister mit den Werten 1 bis #Konstanten. Anschließend werden die Registeranweisungen nacheinander abgearbeitet und der Zustand der Register entsprechend verändert. Das Ergebnis wird zurückgegeben und gespeichert.
Wenn die Karte komplett ausgewertet ist, wird die Summe über alle Rasterzellen berechnet und alle Ergebnisse auf diese Summe normiert, so dass eine Wahrscheinlichkeitsverteilung entsteht. Danach werden Minimum und Maximum dieser Verteilung bestimmt. Für jeden Stichprobenpunkt wird die entsprechende Rasterzelle gesucht und die Wahrscheinlichkeit abgelesen. Diese wird auf das Intervall [0,1] normiert. Ist der normierte Wert größer als eine festgelegte Schwelle, so wird der Hit-Zähler erhöht.
Die Entropie, Mittelwerte und Varianzen werden berechnet und die Differenz zu den Optimalwerten bestimmt. Alle ermittelten Fitnesskomponenten werden gespeichert und mit der Zielfunktion zur Fitness verrechnet.

Zufallsgenerator

Während der Optimierung werden Zufallszahlen benötigt, um zufällig Individuen aus der Population zu ziehen und die Positionen an denen Variationsoperatoren ansetzen auszuwählen. HabitatGP verwendet die GNU scientific library (GSL)[2], welche 13 verschiedene Pseudo-Zufallszahlengeneratoren anbietet. Aus den angebotenen Generatoren wurde `gsl_rng_ranlux389` ausgewählt, da dieser 24 dekorrelierte

Bits [2] verwendet und eine Periode von 10^{171} hat. Diese Eigenschaften sind für Anwendungen mit vielen Auswertungen des Zufallsgenerators (wie GP) vorteilhaft, da Wiederholungen in der Abfolge der Pseudo-Zufallszahlen seltener auftreten. Der Zufallsgenerator wird mit `time(NULL)` initialisiert, der aktuellen Zeit. Dadurch wird jeder Lauf mit einer anderen Folge von Pseudozufallszahlen durchgeführt. Dies ist eine wünschenswerte Eigenschaft, da durch wiederholte Anwendung des Programms ein größerer Bereich des Suchraums abgedeckt wird.

Ausgabe der Ergebnisse

Während der Optimierung wird der Verlauf in Textdateien dokumentiert. Der Zustand der aktuellen Population und die Werte der Fitnesskomponenten jedes Individuums werden in der Datei `fitness.txt` gespeichert, so dass nachträglich die Dynamik in der Änderung der Population nachvollzogen werden kann. In `fitness.txt` werden folgende Informationen (in dieser Reihenfolge) gespeichert: Generation, Anzahl Hits, Fitness, Entropie der Verteilung, summierte Abweichung von Mittelwerten und Varianzen. Darauf folgen für jede Variable erst die Abweichung von jedem Mittelwert in separaten Spalten und danach die Abweichungen der Varianzen in separaten Spalten. In der Datei `statistik.txt` werden für jede Generation die Kenndaten gespeichert. Das sind die Populationsgröße, die Hits des besten Individuums, die Fitness des besten Individuums, die Genomgröße des besten Individuums, die effektive Genomgröße des besten Individuums, die Hits des bisher bestes Individuums, dessen Fitness, Genomgröße und die durchschnittliche Fitness in der Population. Die jeweils besten gefundenen Individuen werden in der Datei `verlauf.txt` dokumentiert. Das HabitatGP kann mehrere Replikate nacheinander berechnen. Der mittlere Fitnessverlauf über die Generationen wird in `meanfit.txt` zusammen mit dem Standardfehler für die mittlere Fitness gespeichert. `laeufe.txt` erhält eine Übersicht über die Fitness des besten gefundenen Individuums jedes Replikats.

3 Eichen-Hainbuchen-Wälder

3.1 Ökologische Motivation

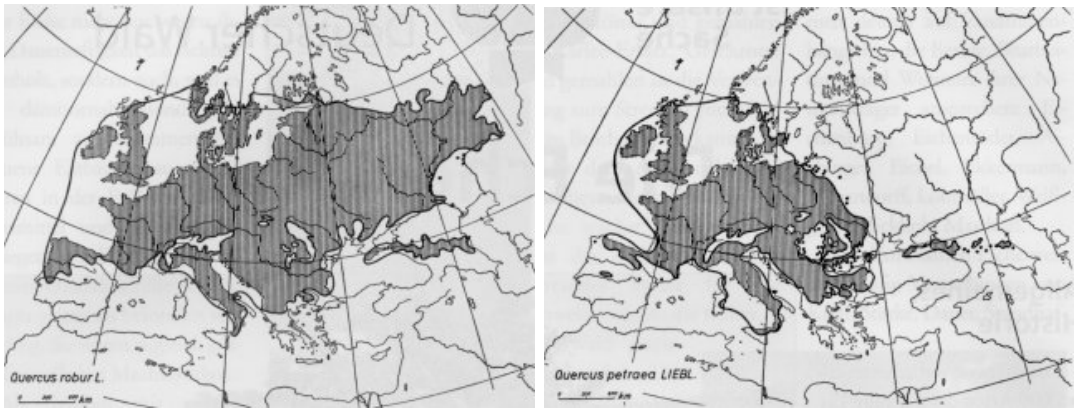
Waldgesellschaften stellen einen Großteil der natürlichen Vegetation Europas dar [16]. Das bisherige Wissen über die Verbreitung von Waldgesellschaften beruht auf Erfahrungswerten und Expertenwissen. In einer europaweiten Kooperation, unter der Leitung von Bohn, wurde dieses Wissen in einer Karte der potentiellen natürlichen Vegetation (PNV) Europas [7] kondensiert und digital verfügbar gemacht. Die PNV-Karte ist bisher die umfassendste Sammlung der potentiellen Verbreitungsgebiete von Pflanzengesellschaften und bietet die Möglichkeit, die potentiellen Standorte von Waldgesellschaften mit ebenfalls digital verfügbaren Umweltdaten zu verknüpfen und zu charakterisieren. Als Fallbeispiel sollen hier Eichen-Hainbuchen-Mischwälder (genauer der Verband: Carpinion betuli) dienen. Die charakteristische Art des Verbandes Carpinion betuli, die Hainbuche, ist ein Birkengewächs. Ihr Verbreitungsgebiet erstreckt sich von Westeuropa bis nach Westrussland. Im Norden noch an den Südspitzen von Großbritannien und Schweden zu finden, reicht das Ausbreitungsgebiet in Südeuropa bis nach Italien und über den Balkan bis nach Griechenland. Außerhalb von Europa gehören noch der nördliche Streifen der Türkei und der Kaukasus zum potentiellen Ausbreitungsgebiet (siehe Abbildung 3.1). Waldgesellschaften des Verbandes Carpinion betuli enthalten immer einen signifikanten Anteil an Stiel- oder Traubeneichen, deren potentielles Verbreitungsgebiet sehr viel größer ist als das der Hainbuche (siehe Abbildung 3.2). Eichen-Hainbuchen-Wälder können sich nur dort ausbilden, wo beide Arten potentiell vorkommen können. Die potentielle Verbreitung des Carpinion betuli hängt deshalb von vielen weiteren Faktoren ab und bietet ein sehr feingliedriges Bild. Die potentiellen natürlichen Habitate sind aus der PVN-Karte ersichtlich [7]. In Europa sind Eichen-Hainbuchen-Wälder selten im natürlichen, oder naturnahen Zustand vorzufinden, sondern stark anthropogen geprägt. Viele dieser Wälder wurden wegen des fruchtbaren Bodens abgeholzt und als Acker genutzt. Noch erhaltene Eichen-Hainbuchen-Wälder wurden von den Menschen in der Vergangenheit verstärkt als

(a) Verbreitungsgebiet der Hainbuche (*Carpinus betulus* und *Carpinus orientalis*) in Europa. Entnommen aus [19]

(b) Typischer Eichen-Hainbuchen-Wald oberhalb von Jena-Winzerla. (Foto: G. Jetschke)

Abbildung 3.1: Hainbuchen

Holzquelle genutzt [7, 6]. Die Artenkomposition der vorhandenen Wälder wurde speziell dadurch beeinflusst, dass durch den Menschen geprägt, indem durch Niederwaldwirtschaft Hainbuchen, Eichen, Linden und Hasel begünstigt wurden, da sie leicht aus dem Stock wieder ausschlagen. Im Beiwerk zur PNV-Karte schreiben die Autoren zum Zustand der naturnahen Bestände: „Vor allem in Mitteleuropa sind naturnahe Bestände von Eichen-Hainbuchen-Wäldern nur noch selten und meist kleinflächig erhalten; sie sollten - soweit nicht bereits geschehen - möglichst als Totalreservate ausgewiesen werden.“ (S. 268, [7]). Eichen-Hainbuchen-Wälder sind inzwischen in die EU-Fauna-Flora-Habitatrichtlinie (FFH) als geschützter Lebensraum aufgenommen worden (Carpinion betuli 9160 und Galio-Carpinetum 9170) [3]. Wie sich die Areale von Gesellschaften unter Klimawandel verändern werden, ist bis jetzt noch nicht sehr gut untersucht. Fischer schreibt in seinem Buch (S. 31, [16]): „Unsere Vorstellungen über diesbezügliche Arealveränderungen in der Zukunft sind derzeit noch sehr wage.“ Seit dieser Ausage im Jahr 2002 haben sich diese Vorstellungen stärker konkretisiert, z.B. durch die Arbeiten von Kölling [32], aber von einem klaren Bild für alle Arten oder Gesellschaften ist der Wissenstand noch entfernt. Die Modelle sollen Aufschluss darüber geben, unter welchen Bedingungen Eichen-Hainbuchen-Wälder optimale Bedingungen vorfinden, indem das Wissen über die natürliche Verbreitung (PNV-Karte) mit Klimavariablen verknüpft wird. Die gelernten Modelle werden in der vorliegenden Arbeit dann auf ein Klimaszenario für das Jahr 2080 angewandt, um dadurch die Auswirkungen der Klimaerwärmung auf die potentielle Verbreitung zu charakterisieren.



(a) Stieleiche (*Quercus robur* (b) Traubeneiche (*Quercus petraea*)

Abbildung 3.2: Verbreitungsgebiete der im Verband Carpinion betuli enthaltenen Eichenarten in Europa. Entnommen aus [18].

3.2 Konkrete Modellplanung

In diesem Abschnitt sollen die Schritte der konzeptionellen Modellformulierung konkret für die Modellierung der Eichen-Hainbuchen-Wälder ausgeführt werden. Zur besseren Strukturierung werden die zu grundeliegenden Daten separat in Abschnitt 3.2.2 beschrieben.

3.2.1 Konzeptionelles Modell

Das konzeptionelle Model basiert auf der Zielstellung, Eichen-Hainbuchen-Wälder auf großen Maßstab zu modellieren und generelle Aussagen über die Abhängigkeit von Verteilungen und Umwelt zu treffen. Außerdem soll das Modell geeignet sein, um auf eine Kllimawandelszenario angewandt zu werden. Diese gewünschte Ausrichtung des Modells führt zu der Entscheidung, ein empirisches Modell zu erstellen, das auf Klimadaten beruht (siehe auch Abbildung 3.4).

Pseudo-Equilibrium

Für die Modellierung wird angenommen, dass sich die Art oder Gesellschaft im Gleichgewicht mit der Umwelt befindet und insbesondere sich das Areal nicht mehr verändert. Diese Annahme ist nicht immer zutreffend [21], hat aber den Vorteil, dass das Habitatmodell durch statische Methoden erstellt werden kann. Statische Habitatmodelle stellen in diesem Fällen immer Momentaufnahmen der Beziehung von

Prädiktorvariablen und Response dar. Da die Daten aus der Karte der potentiellen natürlichen Vegetation stammen, kann die Gleichgewichtsannahme hier zutreffen. Die potentiellen natürlichen Habitate basieren auf mehr Informationen, als nur im Feld gemachte Beobachtungen.

Art vs. Gesellschaften

In dieser Arbeit soll es um die Verbreitungsgebiete von Waldgesellschaften gehen. Eine Gesellschaft wird durch die verschiedenen in ihr vorkommenden Arten charakterisiert. Pflanzen treten in der Regel nicht in Monokultur auf, sondern bilden immer Gesellschaften. Jede Art in einer Gesellschaft hat eine ökologische Nische. Die Ausbildung einer bestimmten Gesellschaft ist von Umweltvariablen abhängig. Diese Abhängigkeit wird durch die Ansprüche der in der Gesellschaft bevorzugt vorkommenden Arten vermittelt. Wenn die prägende Art einer bestimmten Gesellschaft zu schlechte Bedingungen vorfindet, bildet sich eine andere Gesellschaft aus.
Die Modelle in dieser Arbeit basieren nicht auf der parallelen Modellierung repräsentativer Arten der Gesellschaft, sondern die Gesellschaft wird als Einheit („Pseudo-Spezies“) betrachtet und durch Nicht-Community-Methoden berechnet.

Untersuchungsgebiet

Das in dieser Arbeit betrachtete Untersuchungsgebiet umfasst alle Gebiete, die geographisch zu Europa gehören. Es erstreckt sich von der europäischen Westküste von Portugal und Spanien bis zum Ural. Die nördliche Ausdehnung reicht bis zum Polarmeer. Der südwestlichste Punkte der Karte hat die Koordinaten 32°45’N 11°45’W, der nordöstlichste Punkt 72°31’54”N 66°46’15”E.
Im Süden umfasst die Karte auch die Mittelmeerküsten Afrikas und Teile der Türkei. Diese Regionen gehören geographisch nicht zu Europa, wurden aber trotzdem einbezogen. Dies hat zum Einen den technischen Grund, dass eine rechteckige Karte leichter handhabbar ist, und zum Anderen könnte es sein, dass die Vegetation sich unter dem Einfluss der Klimaveränderung in diese Gebiete verschiebt. Auf die Qualität und Anwendbarkeit der Modellierungstechniken hat das Einbeziehen dieser Regionen keinen Einfluss.

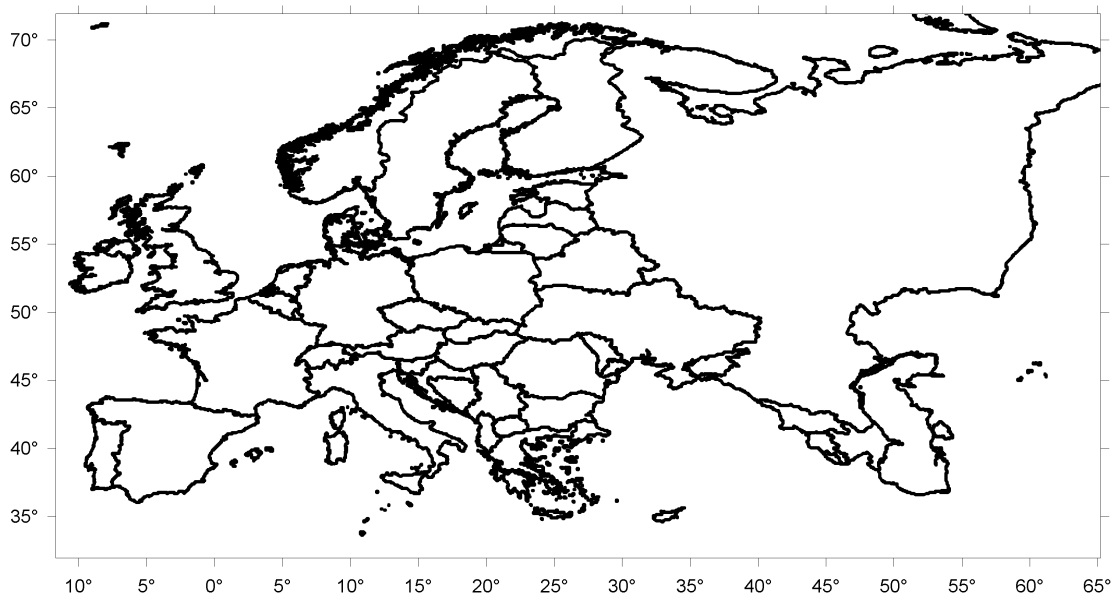


Abbildung 3.3: Karte des Untersuchungsgebiets. Begrenzt durch die Koordinaten 32°45'N 11°45'W und 72°31'54"N 66°46'15"E. In den Rasterkarten sind zusätzlich Gebiete der afrikanischen Mittelmeerküste und Teile der Türkei enthalten.

3.2.2 Datengrundlage

Klimadaten

Für jedes Modell werden Prädiktorvariablen benötigt, aus denen die Habitatgüte berechnet werden soll. Ziel in dieser Arbeit soll es sein, das Vorkommen von Eichen-Hainbuchen-Wäldern in Abhängigkeit von Klimavariablen zu erfassen. Klimavariablen stellen für Pflanzen direkte Prädiktorvariablen dar, im Gegensatz zu indirekten Prädiktoren, wie zum Beispiel Längen- und Breitengrad, die nur mittelbar durch das Klima auf die Pflanzen Einfluss haben. Die Zielsetzung des Modells bestimmt die Wahl der Variablen. In Abbildung 3.4 (aus Guisan und Zimmermann [21]) wird eine Entscheidungshilfe für die Auswahl eines geeigneten Modellansatzes und geeignete Typen von Prädiktorvariablen gegeben.

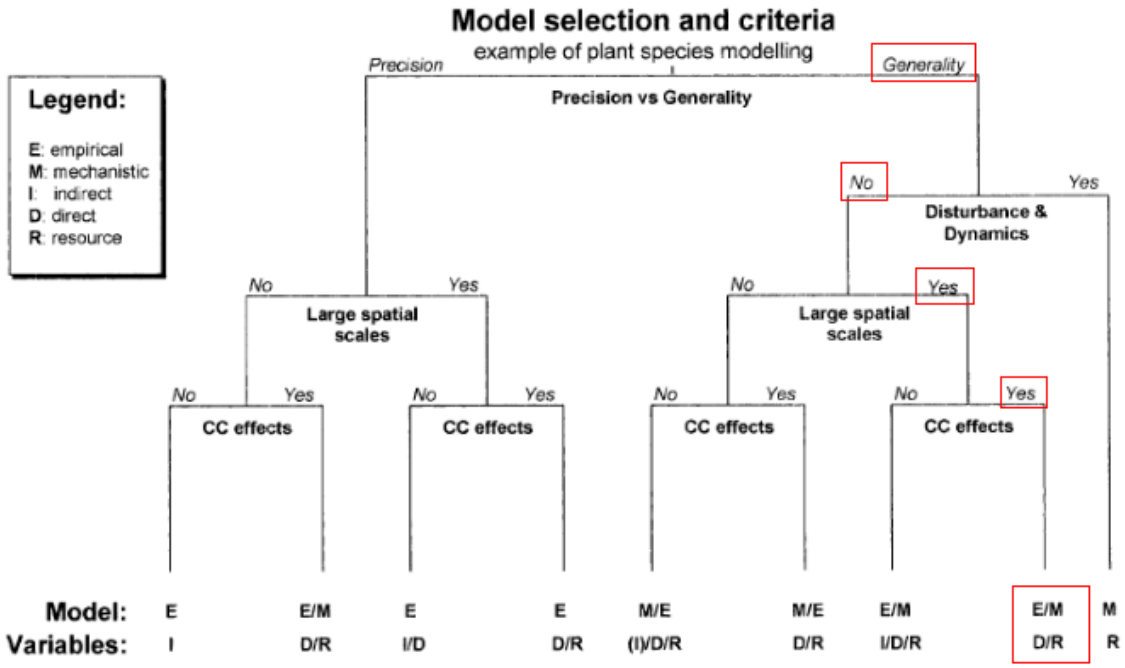


Abbildung 3.4: Entscheidungsbaum: Die Auswahl eines Modelltyps (nach Levins[34]) und geeigneter Prädiktorvariablen wird anhand der Zielsetzung der Modellierung getroffen. Die Zielsetzung dieser Arbeit, ein europaweites Modell der potentiellen Habitate von Eichen-Hainbuchen-Wäldern zu erstellen und auf Klimaszenariodaten anzuwenden, wird durch die markierten Entscheidungen des Baums umrissen. CC: Klimaveränderung. Entnommen aus [21]

Die einrahmten „Entscheidungen“ stellen die Zielsetzung dieser Arbeit dar, ein generelles Modell zu erstellen, welches keine Dynamik berücksichtigt, auf einer großen Skala erstellt wird und im Hinblick auf ein Klimaszenario angewandt wird. Die für diese Ziele geeigneten Modelltypen sind empirische Modelle oder mechanistische Modelle. Geeignete Prädiktorvariablen sind direkte Variablen oder Ressourcen. Im Internet[1] zur Verfügung stehenden Klimadaten stellen für Pflanzen direkte Prädiktoren und Ressourcen (Niederschlag) dar. Die Aufbereitung dieser Klimadaten erfolgte in zwei Schritten [25]. Zuerst wurden die Meßdaten von Wetterstationen über den Zeitraum 1960-1990 gemittelt. Um diese, nur für die nähere Umgebung der Wetterstationen gültigen, Klimadaten für die gesamte Landfläche der Erde bereitzustellen, wurden die Daten interpoliert [25]. Der Interpolationsfehler ist in Europa gering [25], so dass die Modellgüte nicht sehr stark durch fehlerhafte Klimadaten beeinflusst wird. In Tabelle 3.1 (S. 51) sind die verfügbaren Klimadaten zusammengefasst, darunter 36 Basisvariablen, je 12 für die minimale Temperatur, die maximale Temperatur und den Niederschlag, sowie 19 abgeleitete Variablen (Bio-

[1] http://www.worldclim.org

Variablen) [25]. Die Rasterdateien der Umweltdaten sind in einem für geographische

Primäre Prädiktorvariablen		
Name	Beschreibung	Einheit
tmax1-12	maximale Temperatur in den Monaten 1-12	°C · 10
tmin1-12	minimale Temperatur in den Monaten	°C · 10
prec1-12	Niederschlag in den Monaten 1-12	mm
Abgeleitete Prädiktorvariablen		
bio1	Jahresdurchschnittstemperatur	°C · 10
bio2	mittlere monatliche Temperaturdifferenz	°C · 10
bio3	Isothermalität	-
bio4	Temperatursaisonalität (Standardabweichung *100)	°C · 10
bio5	Maximaltemperatur des wärmsten Monats	°C · 10
bio6	Minimaltemperatur des kältesten Monats	°C · 10
bio7	jährliche Temperaturspannweite	°C · 10
bio8	mittlere Temperatur des feuchtesten Quartals	°C · 10
bio9	mittlere Temperatur des trockensten Quartals	°C · 10
bio10	mittlere Temperatur des wärmsten Quartals	°C · 10
bio11	mittlere Temperatur des kältesten Quartals	°C · 10
bio12	jährlicher Niederschlag	mm
bio13	Niederschlag des feuchtesten Monats	mm
bio14	Niederschlag des trockensten Monats	mm
bio15	Niederschlag Seasonalität (Variationskoeffizient)	mm
bio16	Niederschlag des feuchtesten Quartals	mm
bio17	Niederschlag des trockensten Quartals	mm
bio18	Niederschlag des wärmsten Quartals	mm
bio19	Niederschlag des kältesten Quartals	mm

Tabelle 3.1: Verfügbare Umweltvariablen (Quelle: http://www.worldclim.org)

Informations-Systeme (GIS) kompatiblen Format gespeichert und stehen in verschiedenen Auflösungen (Rastergrößen) zur Verfügung (30 Bogensekunden (≈ 1km) bis 10 Bogenminuten (≈ 20km)). In dieser Arbeit werden Klimadaten mit einer Auflösung von 2,5 Bogenminuten (ca. 5km) verwendet. Da die Rasterkarten Daten für die gesamte Erde beinhalten, wurden sie vorher mittels der freien GIS-Software ILWIS auf das Untersuchungsgebiet zugeschnitten.

Für die Projektion der Modelle ins Jahr 2080 werden Daten aus Klimasimulationen verwendet. Verschiedene Einrichtungen haben auf Basis der IPCC-Szenarien

(Intergouvernmental Panel on Climate Change [2]) Klimamodelle entwickelt und durch Simulationen die Entwicklung wichtiger Klimavariablen berechnet. Die Ergebnisse der Simulationen wurden für die Basisvariablen Temperatur und Niederschlag auf worldclim.org aufbereitet und kompatibel gemacht. Die abgeleiteten Variablen (Bio-Variablen) sind noch nicht verfügbar. Die verwendeten Simulationsdaten stammen aus dem CCCMA-Modell (Canadian Centre for Climate Modelling and Analysis) des IPCC-Szenarios A2a. Die A2-Szenarien gehen für die Simulation des Klimas von einer heterogenen Welt mit einer kontinuierlich anwachsenden Populationsgröße aus. Die Wirtschaft ist regional orientiert, fragmentierter und wächst langsamer als in anderen Szenarien (z.B. A1). Weiterführende Informationen finden sich in der Online-Version des „Special Report on Emmision Scenarios“.[3]

Karte der potentiellen natürlichen Vegetation

Als Grundlage für die computerbasierte Stichprobenerhebung dient die Karte der potentiellen natürlichen Vegetation von Bohn et al. [7]. Diese Karte ist sowohl als Druckwerk als auch als digitale Polygonkarte (Shapefile) verfügbar. Jedes Polygon repräsentiert eine Pflanzengesellschaft. Das Areal der Gesellschaft wird durch alle Polygone des gleichen Typs repräsentiert. Zur Identifizierung verschiedener Gesellschaften dient ein Code aus einem Großbuchstaben und einer Zahl. Der Buchstabe identifiziert die Formation, während die Zahlen, fortlaufend innerhalb einer Formation, die verschiedenen Ausprägungen erfassen. Eichen-Hainbuchen-Mischwälder gehören zu den mesophytischen, sommergrünen Laubwäldern und fallen damit in die Formation F. Innerhalb dieser Formation belegen sie die Polygone mit den Codes F34 bis F69 (Carpinion betuli). Die Codes F34 bis F49 enthalten die Stieleichen-Hainbuchen-Gesellschaften, während F50 bis F69 Traubeneichen-Hainbuchen-Wälder enthalten. Das Hauptverbreitungsgebiet (siehe auch Abschnitt 3.1) liegt in und östlich von Polen, aber auch in Deutschland, der Po-Ebene und Frankreich ([7], Abbildung 3.5).

[2] http://www.ipcc.ch
[3] http://www.grida.no/climate/ipcc/emission/

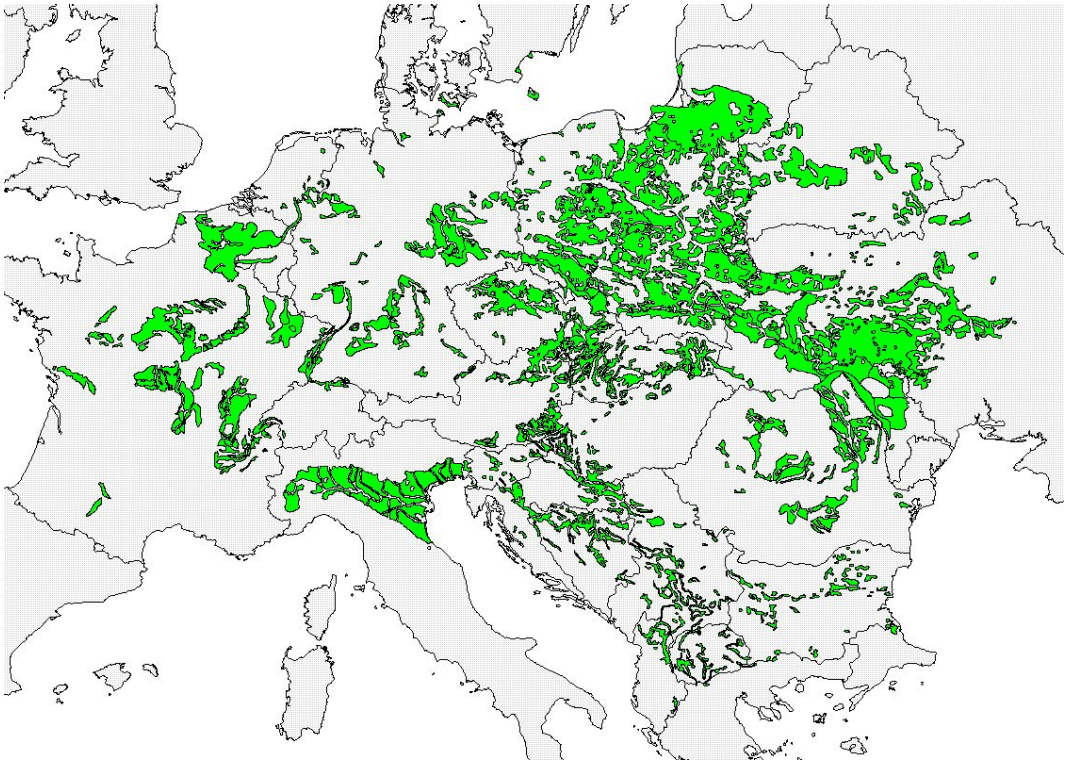


Abbildung 3.5: Ausschnitt der PNV-Karte [7]. Dargestellt sind die potentiellen natürlichen Habitate von Eichen-Hainbuchen-Wäldern (F34-F69).

Insbesondere sollen Modelle speziell für die Vegetationseinheiten F50-F59, die mitteleuropäischen Waldlabkraut-Hainbuchen-Wälder (Galio-Carpinetum), die auch in Deutschland vorkommen, und deren Teilgesellschaften F50-F54, F55-F57, F58-F59 erstellt werden. Zum Einen sollen diese Modelle als Basis für die Evaluierung in Abschnitt 4.1 sein (F50-F59). Zum Anderen soll durch die zusätzlichen Modellierung der Teilgesellschaften überprüft werden, wie konsistent die Teilmodelle F50-F54, F55-F57, F58-F59 mit dem größeren Modell F50-F59 sind.

Stichprobenerhebung

Die Stichprobe wurden aus der durch die PNV-Karte beschriebenen Verbreitung gewonnen. Dazu wurden jeweils die entsprechenden Vegetationseinheiten in der GIS-Software ArcView anhand des PNV-Codes ausgewählt und ein flächenproportionales Sampling angewandt. Aus alle ausgewählten Polygonen sollen N Punkte zufällig ausgewählt werden. Für jedes Polygon i berechnet sich die Anzahl der zu ermittelnden Punkte aus diesem Polygon durch

$$s_i = N \cdot Fläche(i)/Gesamtfläche. \tag{3.1}$$

Da nur ganzahlige Anzahlen von Punkten erhoben werden können, wird s_i auf ganze Zahlen abgerundet. Die ArcView-Erweiterung „Random Point Generator v1.4“ wurde verwendet, um die s_i Punkte zufällig aus den ausgewählten Polygonen zu ermitteln. Durch die Abrundung auf ganze Zahlen kommt es vor, dass bei kleinen Stichprobengrößen Polygone keine Stichprobenpunkte erhalten. Für die weitere Verwendung der Stichproben wurden die Koordinaten der Stichprobenpunkte vom Gradmaß in das Dezimalformat umgerechnet und in CSV-Dateien (comma separated values) exportiert.

4 Anwendung der Maximum-Entropie-Methode

Für die Erstellung von Habitatmodellen mittels der Maximum-Entropie-Methode steht eine Software, MAXENT [47], zur freien Verfügung. MAXENT benötigt als Eingabe eine Menge von „Environmental Layers“ (Rasterkarten). Hier können die WorldClim-Klimaoberflächen verwendet werden. Weiterhin können eine oder mehrere Präsenzstichproben parallel eingelesen werden. Die Präsenzstichproben werden dazu in einer CSV-Datei gespeichert, welche ebenfalls ein vordefiniertes Format aufweisen sollten: In der ersten Spalte wird für jedes Stichprobenelement die Bezeichung der Gesellschaft angegeben, danach folgen Längen- und Breitengradangaben im Dezimalformat.
MAXENT bietet die Möglichkeit, verschiedene Featuretypen auszuwählen. Gehen die Umweltvariablen ohne Veränderung in die Linearkombination der Gibbs-Verteilung ein, werden diese **lineare Features** (L) genannt. Bei **quadratischen Features** (Q) gehen die Umweltvariablen als Quadrat in die Linearkombination im Exponenten der Gibbs-Verteilung ein. **Produkt-Features** (P) erlauben es zwei Umweltvariablen miteinander zu multiplizieren und dadurch Interaktionen zwischen zwei Umweltvariablen zu modellieren. Als weitere Möglichkeiten werden **Threshold-** und **Hinge-Features** angeboten. Diese können die Stichprobe sehr genau erfassen entsprechen allerdings Unstetigkeiten in der Modellfunktion, daher werden im weiteren Verlauf die LQP-Features verwendet.
Die Ausgabe der Modellergebnisse werden von MAXENT als Html-Datei aufbereitet, in der alle wichtigen Informationen zusammengefasst sind. Zusätzlich werden die Modellergebnisse in den Dateien maxentResults.csv, $< art >$_omission.csv und $< art >$_sampleprediction.csv zur weiteren Evaluation bereitgestellt.

4.1 Evaluierung der Maximum-Entropie-Methode in verschiedenen Szenarien

Für alle Szenarien wurden Modelle der Vegetationseinheiten F50-F59 erstellt. Diese wurden mit 2220 Punkte gesampelt.

Einstellung	**Wert**
Feature Typen	LQP
Ausgabeformat	Logistisch
Regularization multiplier	1
Maximum Iteration	500
Convergence Treshold	10^{-5}
Maximum Backgroundpoints	10000
Random Seed	ja

Tabelle 4.1: Allgemeine Einstellungen für MAXENT [47] für die Läufe aus Abschnitt 4.1

4.1.1 Einfluss der Stichprobengröße

MaxEnt ist ein Ansatz, welcher schon aus wenigen Stichprobenpunkten gute Modelle rechnen kann [49], daher soll hier evaluiert werden, wie sich die Modellgüte bei verschiedenen Stichprobengrößen verhält. Dazu wurde der Parameter *Random Testpercentage* variiert, um die Stichprobe automatisch in eine Lern- und eine Teststichprobe zu unterteilen. Der Parameter wurde für die folgende Läufe mit 90 (10%), 95 (5%) und 99 (1%) belegt. Als Variablensatz wurden die 19 Bio-Variablen verwendet.

Für jedes Stichprobengröße wurden 10 Replikate, also gleiche Läufe mit zufällig ausgewählten Stichprobenpunkten, gerechnet. Die zufällige Auswahl der Stichprobenpunkte wurde durch MAXENT während der Laufzeit vorgenommen. Dadurch entspricht diese Versuchanordnung keiner echten Pseudo-Kreuzvalidierung, da die Stichprobe nicht fest partitioniert ist, sondern eher einer Pseudo-Kreuzvalidierung, weil Stichprobenpunkte auch in mehrere Replikate eingehen können. In der Tabelle sind die Mittelwerte $\overline{AUC_{train,test}}$ der Test- und Trainingsstichprobe und die mittlere Anpassung $\overline{R^2}$ aus 10 Replikaten angegeben. Für jedes Replikat wurden die Stichprobenelemente zufällig ausgewählt.

Folgende statistische Auswertungen wurde für die Gruppen 1%, 5% und 10% durch-

geführt. Der Kolmogorov-Smirnov-Anpassungstest zeigt, dass für keine der Gruppen und Gütemaße die Normalverteilungsannahme abgelehnt werden kann (siehe Tabelle 4.2). Der Test der Paardifferenzen ($AUC_{train} - AUC_{test}$) hat die Normalverteilungsannahme ebenfalls nicht abgelehnt ($p_{G1} = 0,755, p_{G5} = 0,329, p_{G10} = 0,303$). Da somit für alle Gruppen die Normalverteilung angenommen werden kann, werden für die folgende Untersuchung eine einfaktorielle ANOVA und der paarige t-Test verwendet. Zusätzlich werden die mittleren PE-Verläufe der verschiedenen Stichprobengrößen dargestellt. Das volle Modell wurde hier nicht betrachtet, da keine Kreuzvalidierungdaten vorhanden sind.

	p-Werte für		
Stichprobengröße	R^2	AUC_{train}	AUC_{test}
1%	0,753	0,688	0,792
5%	0,739	0,394	0,710
10%	0,998	0,900	0,211

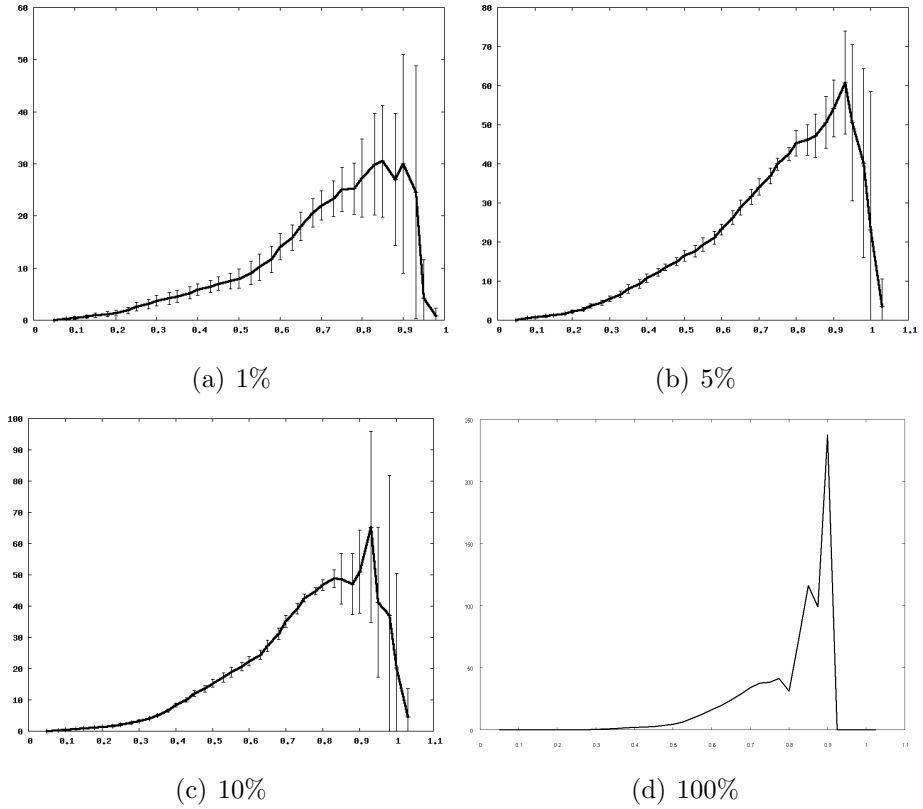
Tabelle 4.2: p-Werte des Kolmogorov-Smirnov-Anpassungstests der Modellgüten für unterschiedliche Stichprobengrößen. Für alle drei Gütemaße (siehe Abschnitt 2.2.5) kann die Normalverteilungsannahme beibehalten werden.

Anzahl Lernstich-proben-punkte	Anteil Lernstich-probe	$\overline{R^2}$	$\overline{AUC_{train}}$	$\overline{AUC_{test}}$	p-Wert
23	1%	0,2793	0,970 (0,003)	0,951 (0,003)	$p < 0,05$
111	5%	0,2157	0,982 (0,0008)	0,977 (0,0005)	$p < 0,0002$
220	10%	0,2545	0,981 (0,0003)	0,979 (0,0003)	$p < 0,0004$
2220	100%	0,47	0,981	-	-

Tabelle 4.3: Mittlere Modellgüte bei verschiedenen Stichprobengrößen. Für jede Stichprobengröße wurden 10 Modelle mit zufälliger Lernstichprobe gerechnet. Die p-Werte entsprechen der Signifikanz des paarigen t-Tests mit der Nullhypothese, dass AUC_{train} sich im Mittel nicht von AUC_{test} unterscheidet. Für alle drei getesteten Gruppen zeigt sich ein signifikanten Unterschied in den AUC-Werten.

Die PE-Graphen (Abbildung 4.1) zeigen deutliche Unterschiede zwischen den vier Stichprobengrößen. Der Kurvenverlauf für die 5%-Modelle und 10%-Modelle entspricht guten Modellen. Die Konfidenzintervalle sind bei den 5%-Modellen kleiner,

dafür scheinen die 10%-Modelle ein höheres Potential in Bezug auf die absolute Größe der PE-Ratio zu besitzen. Das 1%-Modell und das volle Modell zeigen schlechte PE-Verläufe (siehe Beschreibung von PE-Graphen in Kapitel 2.2.5).



(a) 1% (b) 5%

(c) 10% (d) 100%

Abbildung 4.1: mittlere PE-Ratio (siehe Kapitel 2.2.5) der vier Modelle mit 95%-Konvidenzintervall über 10 Replikate

Die Anwendung der einfaktoriellen ANOVA auf die Werte für AUC und R^2 zeigt, dass sich sowohl AUC_{train} ($p < 1 \times 10^{-4}$), AUC_{test} ($p < 1 \times 10^{-10}$) als auch die Anpassung ($p < 0,05$) im Mittel in mindestens zwei der drei Gruppen (1%,5%,10%) unterscheiden. Der Post-Hoc-Test für AUC_{train} und der Post-Hoc-Test für AUC_{test} zeigen, dass die 5%- und 10%-Modelle eine homogene Untergruppe bilden und sich signifikant von den 1%-Modellen unterscheiden. Bei R^2 unterscheiden sich 1% und 5% signifikant, während die 10%-Gruppe sich nicht signifikant zu den anderen Beiden unterscheidet. Die Modelle, die aus 10% der Stichprobe berechnet wurden, haben eine bessere Modellgüte zur Folge. Da der mittlere PE-Verlauf für die 10%-Stichprobe ein gutes Modell anzeigt, die Konfidenzintervalle andeuten, dass die PE-Ratio größer ist als bei 5%-Modellen und die restliche Statistik für 10% eine signifikante Verbesserung im Vergleich zu den anderen Stichprobengrößen anzeigt, werden die weiteren Testläufe mit 10% der Stichprobe gerechnet.

4.1.2 Einfluss korrelierter Variablen

Hier soll charakterisiert werden, inwiefern die Modellgüte der Maxent-Modelle von der Korrelation der Variablen abhängt. Dazu wurde die Umwelt für alle Stichprobenpunkte bestimmt und die lineare Korrelation (nach Pearson) aller Variablen berechnet. Aus diesen Werten wurden mehrere Paare von stark positiv, stark negativ und nicht korrelierten Variablen ausgewählt. Für jedes Paar wurden Modelle gerechnet und die Güte bestimmt. Die Modelle wurden mit 220 Lernstichprobenpunkten berechnet. Sowohl die ausgewählten Paare als auch die Güte der Modelle sind in Tabelle 4.4 angegeben.

Name	V 1	V2	r	$\overline{R^2}$	$\overline{AUC_{train}}$	$\overline{AUC_{test}}$	p-Wert
K+ 1	bio3	tmax2	0,928	0,11	0,931 (0,001)	0,930 (0,001)	$p = 0,473$
K+ 2	bio19	prec12	0,988	-0,03	0,694 (0,004)	0,694 (0,001)	$p = 0,891$
K+ 3	bio10	tmax6	0,954	-0,03	0,903 (0,002)	0,902 (0,0008)	$p = 0,538$
K0 1	prec11	tmin6	0,022	-0,01	0,890 (0,002)	0,889 (0,001)	$p = 0,448$
K0 2	bio13	tmax7	0,005	0,01	0,868 (0,002)	0,862 (0,001)	$p < 0,05$
K0 3	bio12	bio10	-0,006	0,01	0,902 (0,001)	0,901 (0,0004)	$p = 0,43$
K0 4	prec5	tmax5	-0,009	0,04	0,933 (0,001)	0,935 (0,0006)	$p = 0,064$
K- 1	tmax1	bio4	-0,888	0,17	0,953 (0,0007)	0,952 (0,0007)	$p = 0,415$
K- 2	bio15	bio17	-0,825	-0,08	0,787 (0,003)	0,783 (0,002)	$p = 0,247$
K- 3	bio4	tmin1	-0,926	0,14	0,945 (0,0008)	0,888 (0,0542)	$p = 0,332$

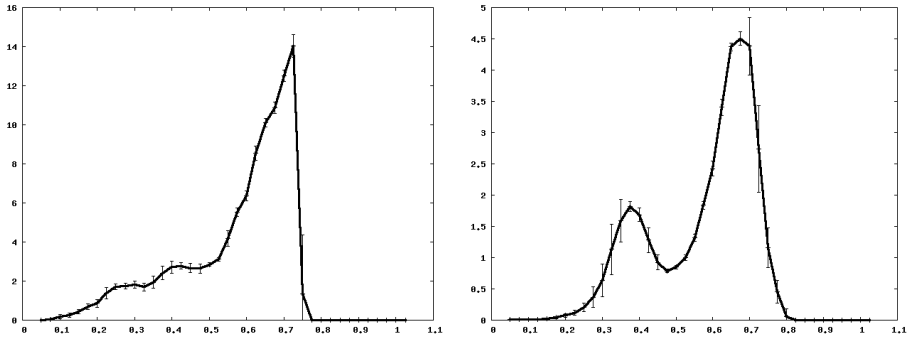
Tabelle 4.4: Mittlere Güte der Modelle aus der Korrelationsuntersuchung (je 10 Replikate). r ist der Pearsonsche Korrelationskoeffizient und gibt die Stärke eines linearen Zusammenhangs der Variablen wieder. Der p-Wert gibt die Signifikanz für den paarigen t-Test von $\overline{AUC_{train}}$ und $\overline{AUC_{test}}$ an. Nullhypothese: $\overline{AUC_{train}} = \overline{AUC_{test}}$.

Die Gütemaße in Tabelle 4.4 lassen auf den ersten Blick keinen Unterschied zwischen den Gruppen (K0, K+, K-) erkennen. Der AUC_{train} liegt im Mittel (über alle Versuche und Gruppen) bei 0,888 (Standardfehler: 0,021) und R^2 bei 0,12 (Standardfehler: 0,03). Die Gruppenmittelwerte von AUC_{train} sind: 0,89865 (K0), 0,895 (K-) und 0,843 (K+). Der Kolmogorov-Smirnov-Test (KS-Test) zeigt, dass die AUC-Werte von K0 normalverteilt sind, die Normalverteilungsannahme für K+ ($p < 0,0005$) und K- ($p < 0,005$) allerdings nicht gilt. Daher wird hier der nichtparametrische Kruskal-Wallis Test zum Vergleich der drei Gruppenmittelwerte verwendet. Dieser zeigt, dass die Korrelation sowohl den AUC_{train} ($p < 0,005$) als auch den AUC_{test}

($p < 0,05$) beeinflusst. Der Kruskal-Wallis-Test basiert auf Rängen. Die mittleren Ränge für jede Gruppe legen nahe, dass K0 (45,65 bzw. 46,94) und K+ (42,02 bzw. 42,60) eine homogene Untergruppe bilden und sich von K- (65,45 bzw. 66,7) unterscheiden. Dieses Ergebnis entspricht nicht der Erwartung, dass ein Unterschied zwischen vorhandener Korrelation (positiver oder negativer) und fehlender Korrelation besteht.
Für die R^2-Werte von K+ ($p = 0,271$) und K0 ($p = 0,361$) lehnt der KS-Test die Normalverteilungsannahme nicht ab, die R^2-Werte von K- sind mit $p < 0,05$ allerdings nicht normalverteilt, so dass für alle Gruppen der nichtparametrische Kruskal-Wallis-Test verwendet wird. Mit $p < 0,01$ zeigt der Test Unterschiede zwischen den Gruppen an. Die mittleren Ränge legen die Vermutung nahe, dass K+ (40,43) und K0 (45,9) eine homogene Untergruppe bilden, während K- (66,7) wahrscheinlich nicht zu dieser Gruppe gehört.
Die ANOVA zeichnet für alle Gütemaße ein einheitliches Bild. Modelle auf der Basis positiver und nicht korrelierter Variablen werden ähnlich bewertet, während negativ korrelierte Variablen einen Einfluss auf die Modellgüte haben. Die Modellgüten sind in jedem Fall stark von den verwendeten Variablen abhängig. Die Modelle K+1 und K-2 zeigen zum Beispiel beide mäßige AUC-Werte, was vermuten lässt, dass nicht die Korrelation, sondern andere Eigenschaften der Variablen die Modellgüte bestimmen und der gerade beschriebene Unterschied zwischen den Gruppen doch zufällig ist. Dies wird auch durch die PE-Graphen unterstrichen. Die Abbildungen 4.2 und 4.3 zeigen die gemittelte PE-Graphen der 10 Replikate für alle Modelle. Die Abbildungen zeigen für alle Modelle einen ähnlichen Verlauf. Der Graph steigt monoton an und fällt bei circa 0,7 auf 0 ab. Dies scheint ein Effekt zu sein, der auf der geringen Variablenanzahl basiert (vergleiche PE-Graphen aus Abschnitt 4.1.1). Ein Teil der Modelle zeigt nicht einmal einen monotonen Anstieg, sondern eine bimodale Form. Die PE-Graphen von K- unterstreichen den Eindruck, dass K- die besseren Modelle erzeugt, denn diese Graphen weisen alle keine bimodale Form auf und die Werte der PE-Ratio sind im Mittel höher als für K+- und K0-Modelle.

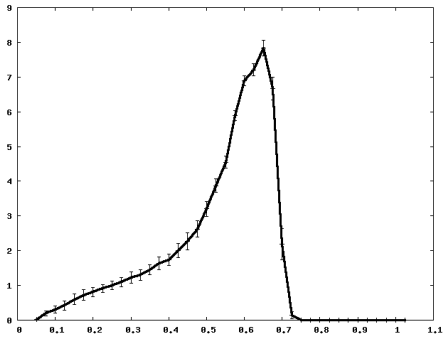
4.1.3 Variablenselektion nach Diskriminanzanalyse

Eine Möglichkeit eine Vorauswahl von Prädiktorvariablen vorzunehmen, ist die Diskriminanzanalyse. Dieses statistische Verfahren bringt eine Menge von Variablen in eine Reihenfolge, indem eine Diskriminanzfunktion iterativ aufgebaut wird. Die Reihenfolge der Variablen basiert auf der durch die Variable erklärte Varianz.
Die durch die Diskriminanzanalyse ausgewählten Variablen dienen als Basis für die Entwicklung des Modells in diesem Abschnitt. Analog zur Reihenfolge werden nach-

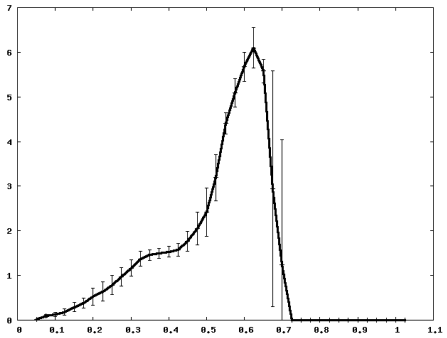


(a) K+1

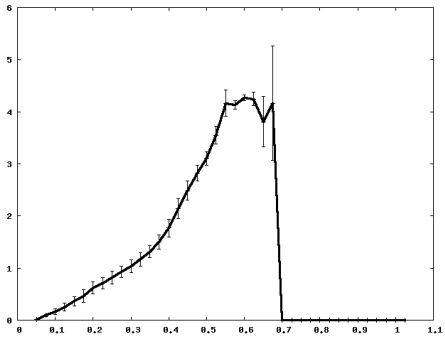
(b) K+2



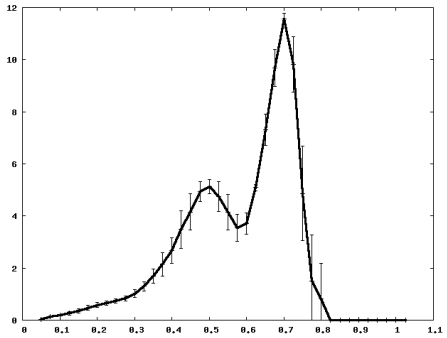
(c) K+3



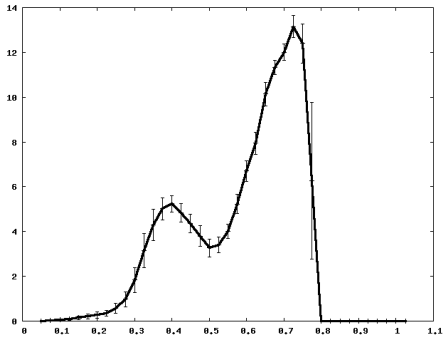
(d) K01



(e) K02

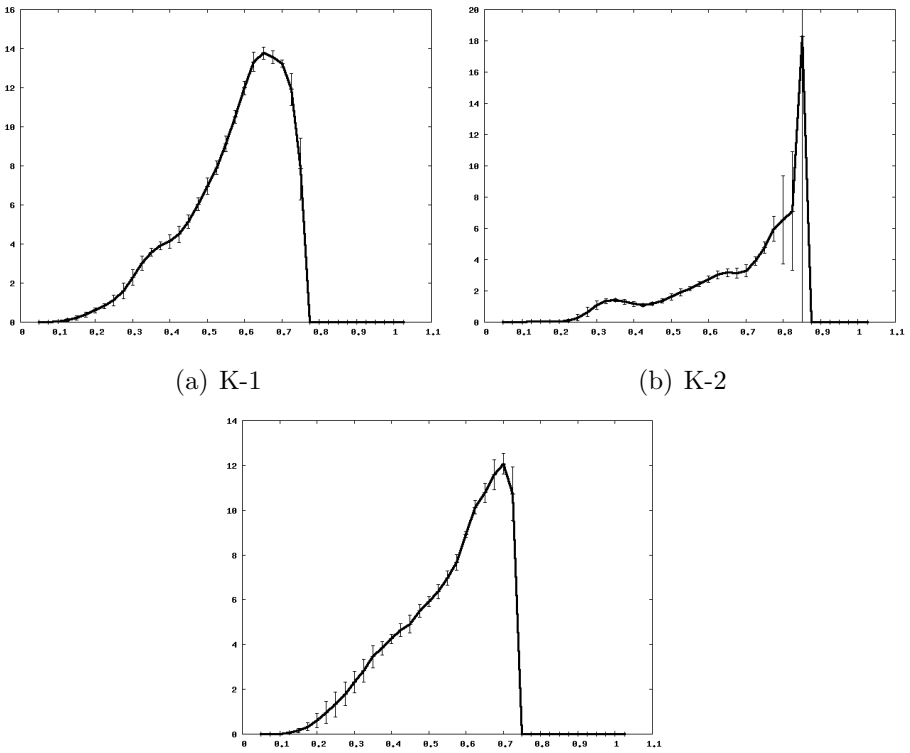


(f) K03



(g) K04

Abbildung 4.2: PE-Ratio für die Modelle auf Basis positiv korrelierter Variablen (K+) und nichtkorrelierter Variablen (K0). Die Graphen zeigen die Mittelwerte über 10 Replikate für die verschiedenen Modelle (siehe Tabelle 4.4). Die Fehlerbalken zeigen das 95%-Konfidenzintervall für jeden Wert. K+ und K0 konnten durch die statistische Auswertung als nicht signifikant unterschiedlich bewertet werden.



(a) K-1 (b) K-2

(c) K-3

Abbildung 4.3: PE-Ratio für die Modelle auf Basis negativ korrelierter Variablen (K-). Die Graphen zeigen die Mittelwerte über 10 Replikate der Modelle aus Tabelle 4.4. Die Fehlerbalken zeigen das 95%-Konfidenzintervall für jeden Wert. In den PE-Graphen sieht man durch den monotoneren Verlauf und absolut größere PE-Ratio die bessere Performanz von (K-)-Modellen im Vergleich zu K+ und K0-Modellen.

einander 10 Modelle erstellt. Jedes Modell enthält alle Variablen des Vorgängers und zusätzlich eine neue Variable. So kann untersucht werden, wie sich die Modellgüte bei der Hinzunahme von Variablen verändert.
Die Anwendung einer Diskriminanzanalyse hat den Nachteil, dass eine Absenzstichprobe benötigt wird. Da die Daten auf Grundlage der PNV-Karte erhoben wurden, enthält die erhobene Absenzstichprobe nur echte Absenzpunkte, so dass die Diskriminanzanalyse nicht durch Pseudo-Absenz verfälscht wird.
Die Ergebnisse sind in Tabelle 4.5 zusammengefasst. Die Abbildungen II.2 bis II.11 im Anhang zeigen die entsprechenden Wahrscheinlichkeitsverteilungen. Anhand der Abbildungen lässt sich erkennen, wie MaxEnt die zusätzlichen Informationen in jedem Schritt verwendet, um die Verteilung anzupassen.
Ein Modell, das so erstellt wurde, sollte eine große Diskriminanz zeigen, da die Variablen nach diesem Kriterium ausgewählt wurden.

Name	Neue Variable	AUC_{train}	AUC_{test}	R^2
D1	tmax3	0,884	0,884	-0.03
D2	tmax11	0,965	0,966	0.04
D3	bio6	0,970	0,971	0.14
D4	bio9	0,970	0,972	0.13
D5	prec10	0,975	0,976	0.11
D6	prec8	0,979	0,979	0.19
D7	tmin11	0,979	0,979	0.17
D8	tmin10	0,980	0,980	0.20
D9	tmax10	0,980	0,980	0.17
D10	tmin4	0,982	0,981	0.20

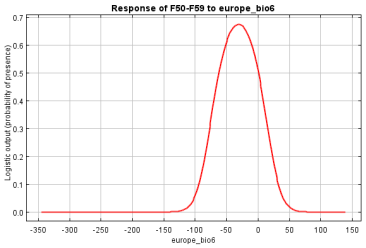
Tabelle 4.5: Güte der Modelle, die aus der Diskriminanzanalyse hervorgegangen sind. Mit jeder neuen Variable ist eine Verbesserung der Anpassung (R^2) und der Diskriminanz (AUC) zu beobachten.

Es lässt sich beobachten, dass die Anpassung mit der Variablenanzahl zuerst zunimmt und dann alterniert. Ein klarer Trend zeigt sich auch beim AUC. Je mehr Variablen hinzukommen, desto besser wird die Diskriminanz. Die Verbesserung bewegt sich allerdings in einem sehr kleinen Bereich, insgesamt eine Verbesserung von 0,098 für AUC_{train} und 0,097 für AUC_{test}. Da die erste Variable schon ein Modell mit einem hohen Wert für AUC erzeugt hat, bringen die neuen Variablen keine neue Informationen, die zur Diskriminierung genutzt werden können, in das Modell ein. MaxEnt schätzt für jedes Modell die für das Modell wichtigste Variable durch den Zuwachs der Modellparameter während der Kalibrierung. Die Variablenwich-

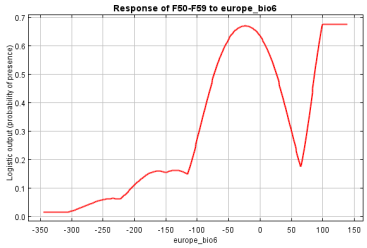
Variable	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
tmax3	100	56,8	18,9	19,2	12,8	9,6	8,8	8,4	10	8,5
tmax11	0	43,2	31,8	12,3	12,8	7,8	6	5,3	5,3	4,7
bio6	0	0	49,3	40	40,1	36,7	36,3	36,8	35,9	35,8
bio9	0	0	0	28,5	26,5	25	25,4	35,1	24,2	24
prec10	0	0	0	0	7,9	9,9	9,4	9,5	8,5	9,1
prec8	0	0	0	0	0	11	9	8,7	8,4	7,6
tmin11	0	0	0	0	0	0	5,1	5,2	6	4,8
tmin10	0	0	0	0	0	0	0	1	0,3	1,9
tmax10	0	0	0	0	0	0	0	0	1,4	0,3
tmin4	0	0	0	0	0	0	0	0	0	3,2

Tabelle 4.6: Prozentualer Anteil der Variablen an den Modellen. Geschätzt durch den Zuwachs der Modellparameter (λ) während der Kalibrierung.

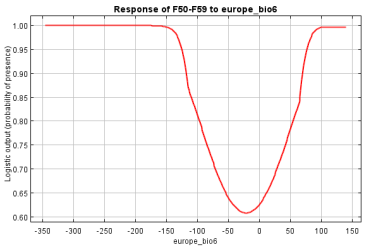
tigkeit wird in Tabelle 4.6 für alle Schritte angegeben. Die Response wird für jede Variable in der Ausgabedatei als Schnitt durch die mehrdimensionale Modellfunktion ausgegeben. Dieser Schnitt wird erzeugt, indem alle anderen Variablen auf ihren empirischen Mittelwert gesetzt werden und die Modellfunktion über betrachtete Variable ausgewertet wird. Die ausgegebene Responsekurven verändern sich mit jedem Modell. So ist bei den meisten Variablen, die ein unimodales Verhalten zeigen, eine Verbreiterung der Responsekurve zu beobachten, je mehr Variablen im Modell enthalten sind. Befindet sich das Maximum der unimodalen Response nahe am Rand des gültigen Bereichs (minimaler Wert und maximaler Wert der Variable), so ändert sich die Response durch diese Verbreiterung zum Beispiel in eine Response, die ein logistisches Verhalten zeigt. Manche Variablen ändern ihr Verhalten komplett. Die Variable bio6 zeigt erst ein unimodales Verhalten und kehrt sich in D9 komplett um, so dass das Minimum an derselben Stelle liegt wie das Maximum vorher. In der Schätzung der Wichtigkeit der Variablen (Percent contribution) steht bio6 mit 35% bis 49% bei jedem Modell dieser Reihe an der Spitze.



(a) Response von bio6 aus D4



(b) Response von bio6 aus D8

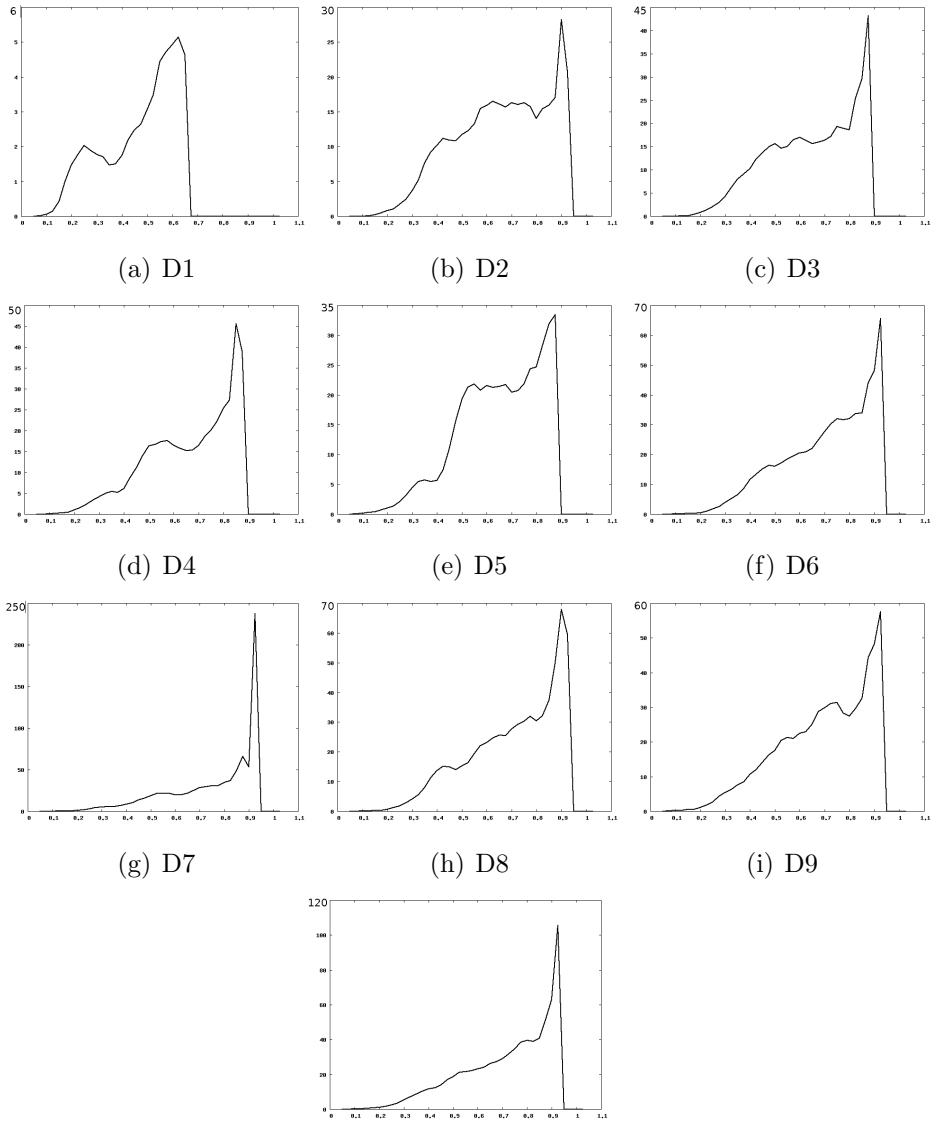


(c) Response von bio6 aus D9

Abbildung 4.4: Response der wichtigsten Modellvariable bio6 (Minimaltemperatur des kältesten Monats). Dargestellt ist die Änderung der Form der Responsekurve. Dadurch lässt sich umreissen, wie MaxEnt die Variablen verwendet, um die Verteilung möglichst gut anzupassen. Eine mögliche Erklärung, wie die unterschiedlichen Funktionsverläufe entstehen wird in Abbildung 4.6 gegeben.

Erklären lässt sich dieses Verhalten durch die Schnitte zur Darstellung der mehrdimensionale Modellfunktion. Mit jeder zusätzlichen Variablen erhält die Modellfunktion eine Dimension mehr. Die neue Achse verändert die Position des Schnittes. Dadurch verschiebt sich auch die Schnittfläche und die Response bekommt eine andere Form, obwohl sich die mehrdimensionale Funktion in der Response dieser Variable nur wenig geändert hat. Abbildung 4.6 zeigt dieses Verhalten an einem Beispiel. Die Abbildungen der Response in der Ausgabe von MaxEnt geben ein unvollständiges Bild der Responsefunktion, da sich auf die durchschnittliche Umwelt beziehen.
Die PE-Graphen der 10 Modelle (Abbildung 4.5) aus diesem Abschnitt zeigen, wie sich die Modellgüte in einzelnen Bereichen der Vorhersage verändert. Die Formen der PE-Graphen scheinen mit R^2 bis zu einem gewissen Grad korreliert zu sein. Die Modelle D5, D7 und D9 stellen alle Modelle dar, in denen R^2 wieder schlechter wurde. Die entsprechenden PE-Graphen weichen stärker von der optimalen Form (monotoner Anstieg bis zum Maximum) ab, als die besseren Modelle D6,D8 und D10, welche näher an einer linearen Form bis zum Maximum sind. Die absolute Höhe schwankt durchgehend. Das Maximum tritt in D7 mit einem Wert in der Nähe von 250 auf. Die Modellgüte von D7 zeigt zwar einen großen Wert für AUC aber kein optimales R^2. Die PE-Kurve legt die Vermutung nahe, dass hohe Variablen im Modell überrepräsentiert sind. Das beste Modell, sowohl in Bezug auf R^2

und PE-Ratio ist D6. Der PE-Graph liegt nahe am Optimum. Die R^2-Werte für Modelle nach D6 alternieren, D6 ist somit das erste Modell mit gutem R^2 und die Diskriminanz ist sehr gut, auch wenn die nachfolgenden Modelle noch höhere Werte für AUC erreichen.



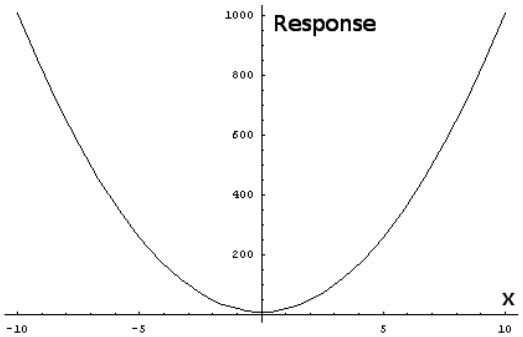
(a) D1 (b) D2 (c) D3

(d) D4 (e) D5 (f) D6

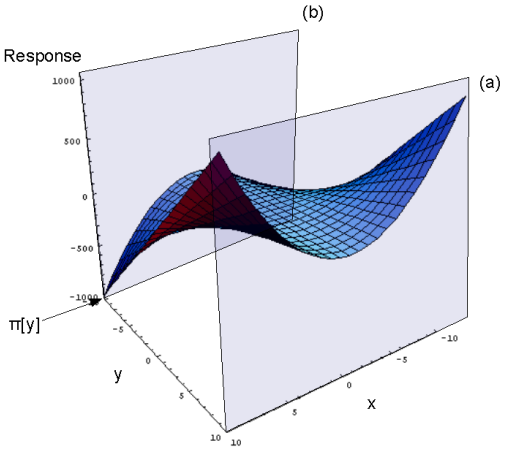
(g) D7 (h) D8 (i) D9

(j) D10

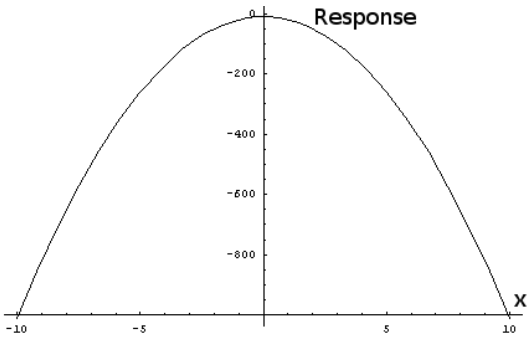
Abbildung 4.5: PE-Graphen der 10 auf der Diskriminanzanalyse basierenden Modelle.



(a) Eindimensionale Response, des Modells mit einer Variable



(b) Zweidimensionale Response. Die zusätzliche Variable habe ihren Mittelwert bei -10



(c) Schnitt durch die zweidimensionale Responsefunktion bei y=-10

Abbildung 4.6: Die Hinzunahme einer Variablen kann den Schnitt durch die mehrdimensionale Modellfunktion und damit die ausgegebene Response einer Variablen verändern und dadurch den Eindruck einer veränderten Response vervorrufen. Beispielfunktion: $Response = x(1+y)^2$.

4.1.4 Zusammenhang von AUC und R^2

Betrachtet man alle Paare von werten für AUC_{train} und R^2, die im Laufe dieser Auswertung aufgetreten sind, so lässt sich auf Grund der gebogenen Form der Punktwolke ein nichtlinearer Zusammenhang vermuten. Bei schlecht angepassten Modelle treten AUC-Werte zwischen 0,65 und 0,99 auf. Schlecht angepasste Modelle haben also nicht unbedingt auch eine schlechte Diskriminanz. Sobald die Anpassung über 0,1 steigt, zeigen die Modelle fast durchgehend ein sehr gute Diskriminierfähigkeit. Eine logistische Regression (Obergrenze bei 1) zeigt mit der Gleichung $\ln\left(\frac{1}{AUC}\right) = 0,122 \cdot 0,003^{R^2}$ ein Bestimmtheitsmaß von 0,618. Dasselbe Bild zeigt sich für AUC_{test} und R^2. Dies lässt sich durch die starke lineare Korrelation ($r = 0,851$) von AUC_{train} und AUC_{test} erklären. Bei den betrachteten Modellen lag R^2 nie über 0,4, so dass über das Verhalten von AUC bei größeren Werten von R^2 keine Aussagen gemacht werden können.

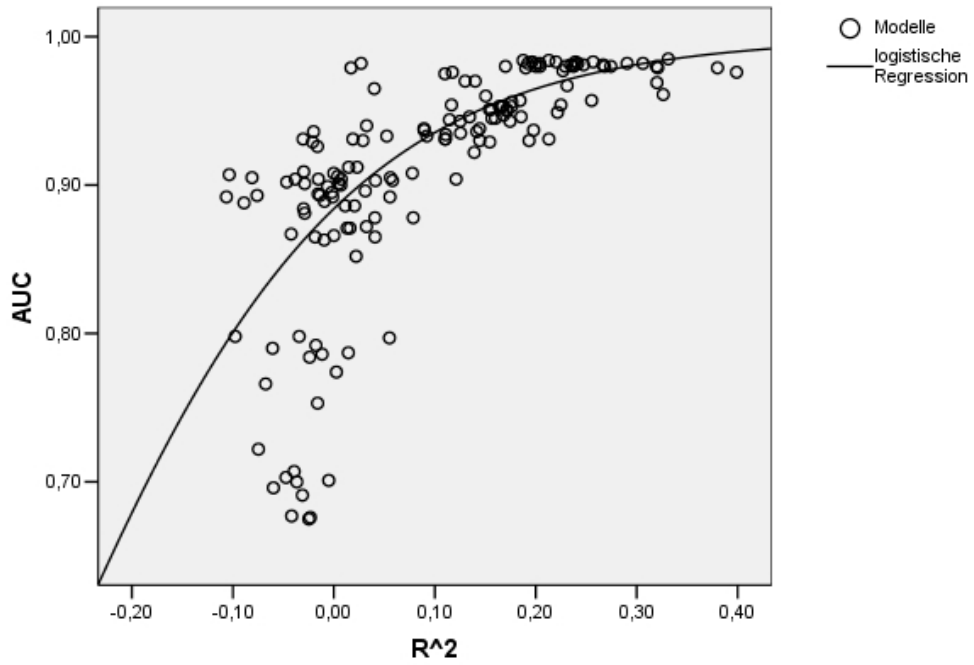


Abbildung 4.7: Streudiagramm aller AUC_{train}-R^2 Wertepaare der Auswertungen aus Abschnitt 4.1. Eingezeichnet ist die Ausgleichskurve einer logistischen Regression ($\ln\left(\frac{1}{AUC}\right) = 0,122 \cdot 0,003^{R^2}$). Für AUC_{test}-R^2 ergibt sich durch die starke Korrelation von AUC_{train} und AUC_{test} ($r = 0,851$) dasselbe Bild.

4.2 Modelle europäischer Eichen-Hainbuchen-Wälder

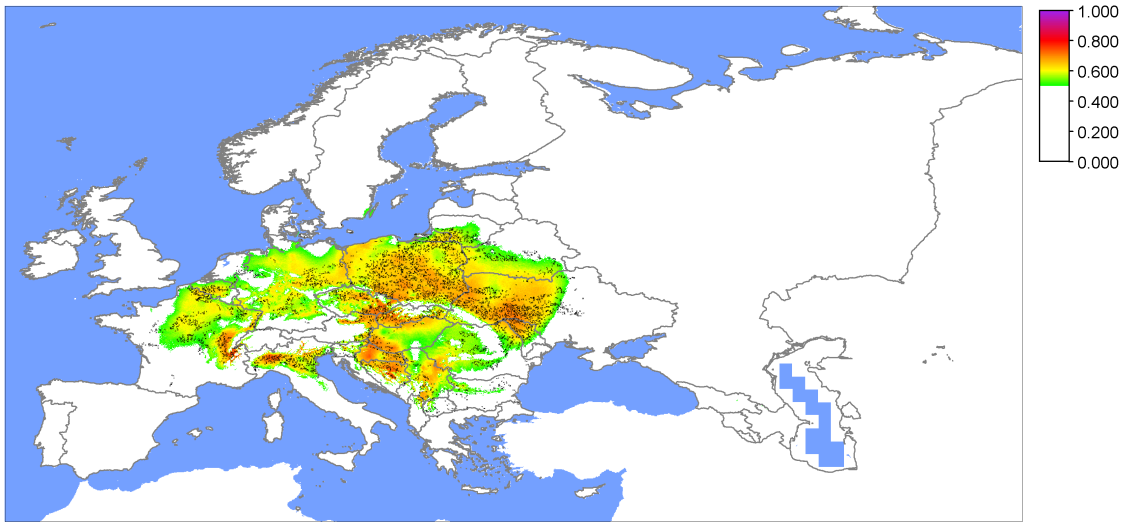
Die verschiedenen hier vorgestellten, Modelle repräsentieren unterschiedlich feine Teilgesellschaften des Verbandes Carpinion. Das Modell mit F34-F69 umfasst die gesamte Ordnung und beschreibt die mittleren Ansprüche aller europäischen Hainbuchen-Wälder. Die drei feinsten Aufgliederungen, F50-F54, F55-F57 und F58-F59, bilden zusammen F50-F59. Auch wenn diese Gesellschaften in den anderen zwei Modellen enthalten sind, so sind doch alle Modelle separat voneinander zu betrachten und zu bewerten. Dass die Modelle hierarchisch ineinandergreifen, muss nicht heißen, dass sie die selben Ergebnisse zeigen bzw. konsistent sind.
Zu jedem Modell sind die Responsekurven der wichtigsten Variablen des Modells wiedergegeben. Ob eine Variable für das Modell wichtig ist, wurde durch MAXENT abgeschätzt, indem während der Berechnung der Modellparameter der prozentuale Anteil der Lambdas dieser Variable am Gesamtmodell berechnet wurde. Eine Übersicht des prozentualen Variablenbeitrags ist für alle Modelle in Abbildung 4.18 auf Seite 82 dargestellt.

Einstellung	**Werte**				
Modellierte Gesellschaften	F34-F69	F50-F59	F50-F54	F55-F57	F58-F59
Größe der Gesamtstichprobe	5615	2220	856	916	448
Random Testpercentage	0				
Verwendte Umweltvariablen:	tmin1-12,tmax1-12,prec1-12				
Feature Typen	LQP				
Ausgabeformat	Logistisch				
Regularization multiplier	1				
Maximum Iteration	500				
Convergence Threshold	10^{-5}				
Maximum Backgroundpoints	10000				
Random Seed	ja				

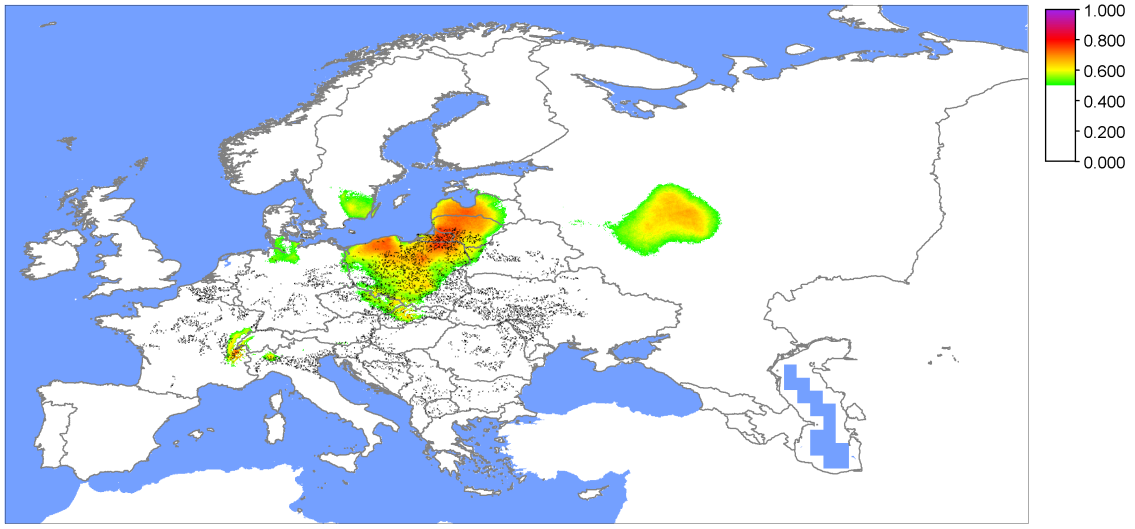
Tabelle 4.7: Einstellungen für MaxEnt für die Läufe aus Abschnitt 4.2

4.2.1 Europäische Eichen-Hainbuchen-Wälder

Die Anwendung des Maximum-Entropie-Modells für die Vegetationseinheiten F34-F69 auf die aktuelle Umwelt ist in Abbildung 4.8 dargestellt. Es lässt sich erkennen,



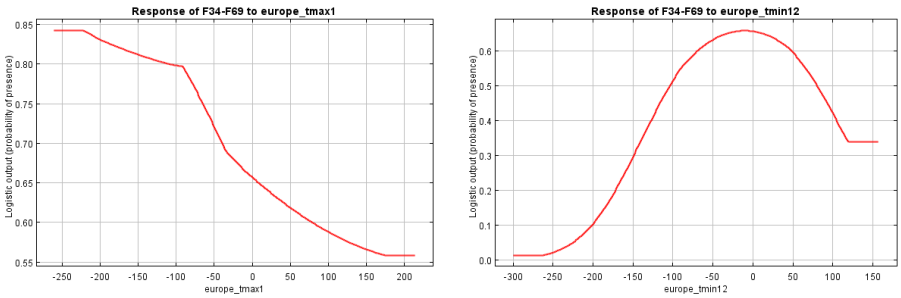
(a) Maximum-Entropie Habitatmodell der europäischen Eichen-Hainbuchen-Wälder. Bestimmung des Kernhabitats durch Anwendung eines Schwellenwerts (0,5).



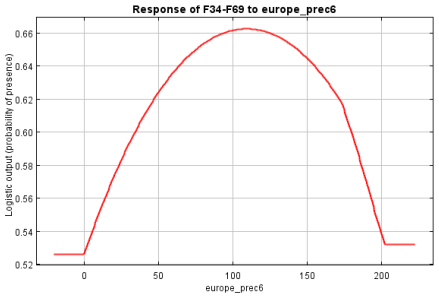
(b) Projektion des MaxEnt-Modells für die Vegetationseinheiten F34-F69 auf simulierte Klimadaten für 2080 (IPCC-Szenario A2a)

Abbildung 4.8: Modellvorhersage für die europäischen Eichen-Hainbuchen-Wälder (F34-F69).

dass das Modell die potentielle natürliche Vegetation gut wiedergibt. Dies wird sowohl durch die Anpassungsgüte (R^2=0,39), als auch die Diskriminanz (AUC=0, 952) ausgedrückt. Das Modell hat gut generalisiert. Die ist an den Rasterzellen zu erkennen, den Wahrscheinlichkeiten über 0,5 zugewiesen wurden, obwohl sie nicht zur Stichprobe gehören. Die Responsekurven (Abbildung 4.9) der drei wichtigsten Variablen des Modells spiegeln die Ansprüche von Eichen-Hainbuchen-Wäldern in Europa wieder. Es zeigt sich, dass die Verteilung hauptsächlich von der Temperatur im Winter anhängt (tmax1 und tmin12) und erst in zweiter Linie von der Regenmenge



(a) tmax1, 30% Modellbeitrag (b) tmin12, 9,6% Modellbeitrag

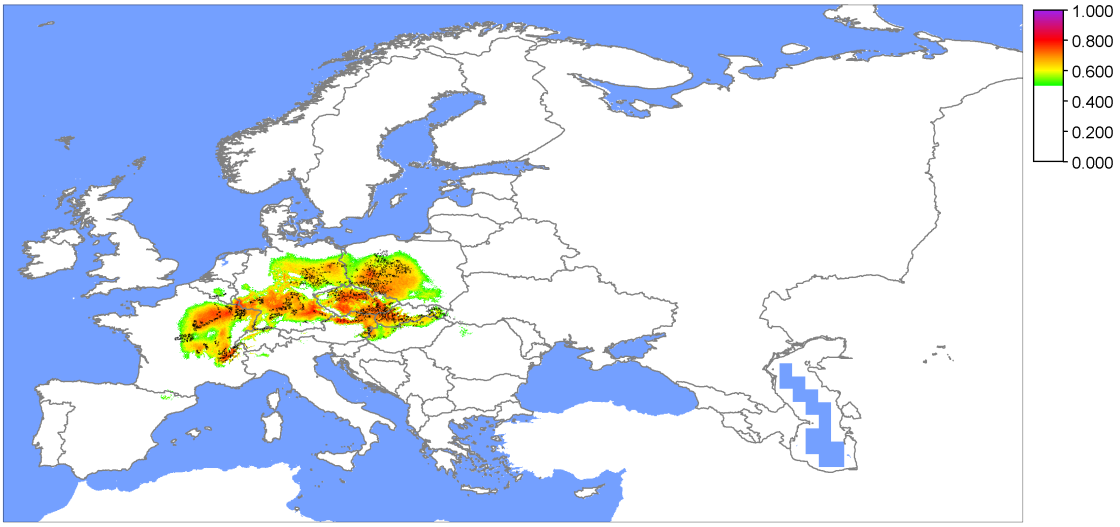


(c) prec6, 6,6% Modellbeitrag

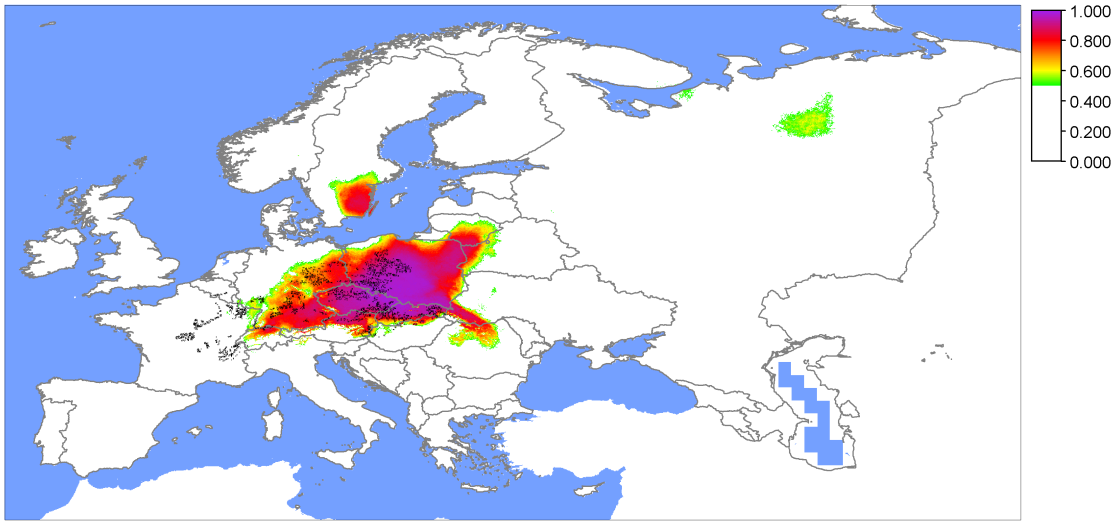
Abbildung 4.9: Responsekurven der wichtigsten Variablen von F34-F69. Temperaturvariablen in °C·10. Niederschlag in mm.

(prec6). So zeigt sich für tmin12, dass bei einer minimalen Dezembertemperatur von circa -2°C das Optimum für den Verband Carpinion betuli liegt. Niedrigere oder auch höhere Temperaturen wirken sich negativ auf die Response aus. Der Niederschlag im Sommer (prec6) zeigt ebenfalls eine unimodale Response mit dem Optimum bei circa 100mm Niederschlag im Juni. Das Modell deckt sich hier mit der Literatur, die für Eichen-Hainbuchen-Wälder den Feuchtebereich mit „trocken bis mäßig trocken“ angibt [6]. Der Niederschlag kann nur einen Hinweis auf die Bodenfeuchte geben, da diese von weiteren Faktoren abhängt. Die Projektion des Modells für die Vegetationseinheiten F34-F69 auf die Umweltdaten für 2080 (Abbildung 4.8) zeigt eine deutliche Verschiebung des potentiellen Habitats nach Nord-Osten. Gleichzeitig verringert sich die Fläche des potentiellen Habitats. Zählt man alle Rasterzellen mit einer Wahrscheinlichkeit größer als 0,5 zum Kernhabitat (0,5 wird dem typischen Habitat zugeordnet [48]), so lässt sich eine Verringerung um 56% von 126865 auf 55360 Rasterzellen beobachten. Dabei verschlechtern sich die Bedingungen in ganz Westeuropa (Frankreich, Italien, Deutschland), aber auch in Ungarn, Rumänien und der Ukraine. In Polen verlagert sich das Kernhabitat in Richtung kurische Nehrung. In dieser Region erweitert sich das Habitat auf Südschweden und die baltischen Staaten, außer Estland.
In der Nähe von Moskau zeigt sich ein weiteres Gebiet mit hohen Wahrscheinlichkeiten, welches 2080 ebenfalls gute Bedingungen für Eichen-Hainbuchen-Wälder bieten könnte.

4.2.2 Mitteleuropäische Eichen-Hainbuchen-Wälder



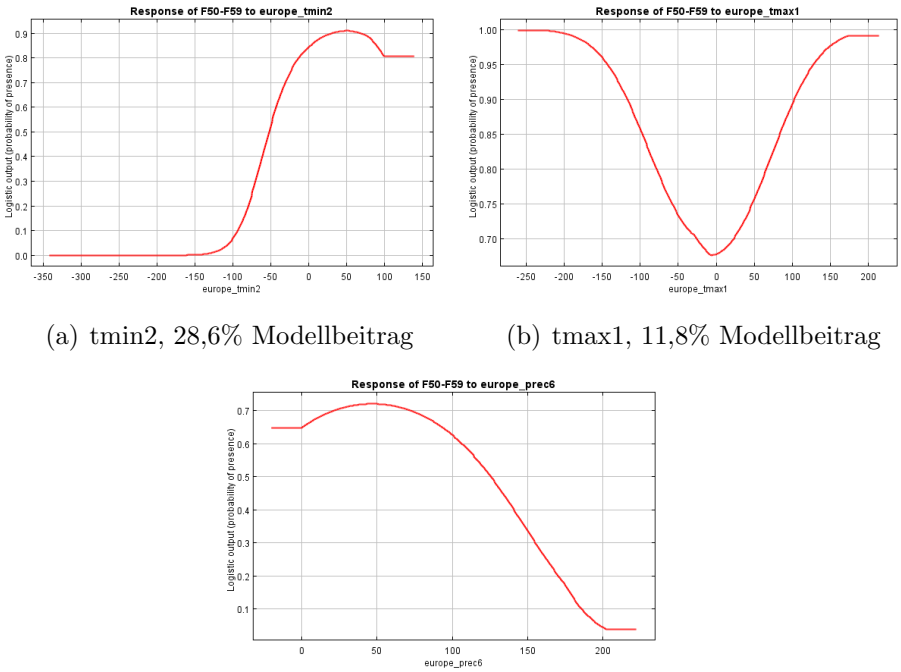
(a) Maximum-Entropie Habitatmodell der mitteleuropäischen Eichen-Hainbuchen-Wälder. Bestimmung des Kernhabitats durch Anwendung eines Schwellenwerts (0,5).



(b) Projektion des MaxEnt-Modells für die Vegetationseinheiten F50-F59 auf simulierte Klimadaten für 2080 (IPCC-Szenario A2a)

Abbildung 4.10: Modellvorhersage für die mitteleuropäischen Eichen-Hainbuchen-Wälder (F50-F59).

Die Vorhersage des Areals für die mitteleuropäischen Eichen-Hainbuchen-Wälder (F50-F59) zeigt eine gute Anpassung an die Stichprobe ($R^2 = 0,51$) und eine gute Diskriminanz ($AUC = 0,995$) zwischen Vorkommen und Nichtvorkommen. Das Modell spiegelt die Vorliebe der Eichen-Hainbuchen-Wälder für die kolline Stufe gut wieder, da zum Beispiel in den Alpen, im Erzgebirge (montane Stufe) und im norddeutschen Tiefland (planare Stufen) die Wahrscheinlichkeiten kleiner als 0,5 sind. Dieses Verhalten kann nicht auf die Höhe direkt zurückgeführt werden, da die



(a) tmin2, 28,6% Modellbeitrag

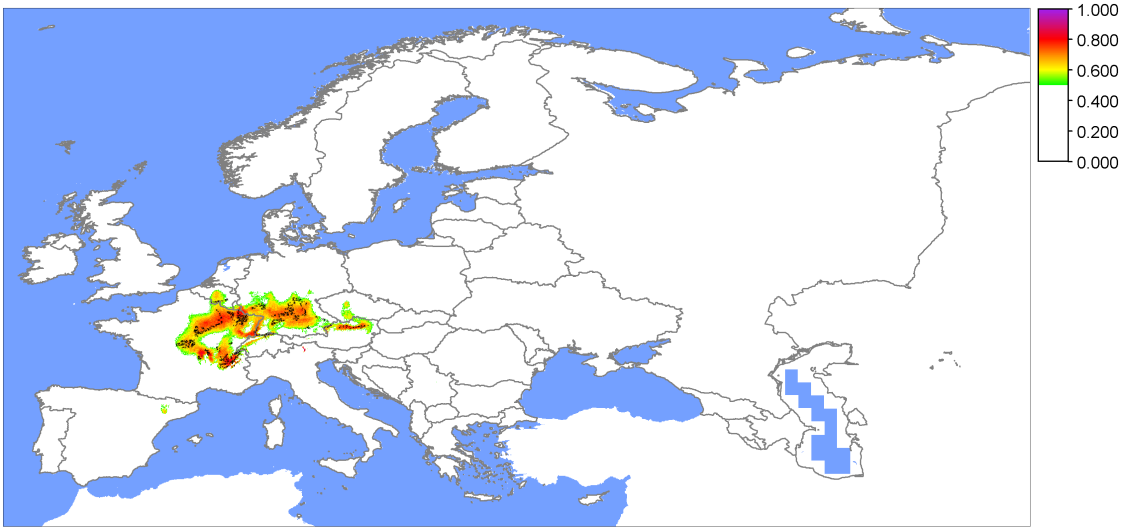
(b) tmax1, 11,8% Modellbeitrag

(c) prec6, 7,5% Modellbeitrag

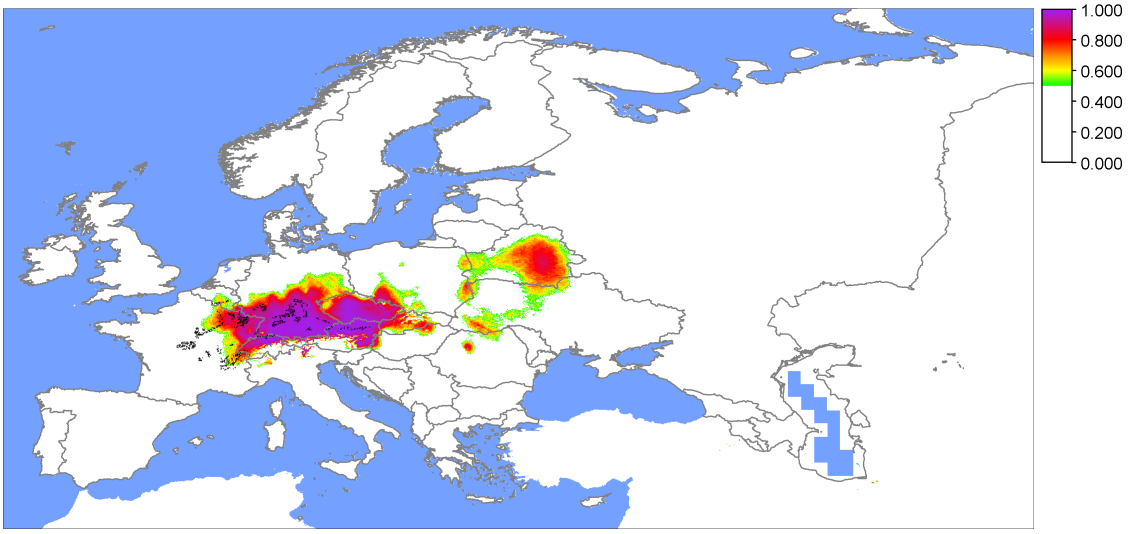
Abbildung 4.11: Responsekurven der wichtigsten Variablen von F50-F59. Temperaturvariablen in °C·10. Niederschlag in mm.

Höhe nicht als Variable in das Modell eingegangen ist, sondern nur auf die Klimavariablen. Für F50-F59 sind tmin2, tmax1, prec6 als Variablen mit dem größten geschätzten Anteil am Modell aufgeführt. Die Verteilung hängt am stärksten von der Temperaturschwankung im Winter und dem Niederschlag im Juni ab (siehe Responsekuven 4.17). tmax1 stellt dabei die obere Grenze für die Temperatur dar. Die Responsekurve zeigt, dass Temperaturen um den Gefrierpunkt sich leicht negativ auf die Waldgesellschaft auswirken. Die minimale Wahrscheinlichkeit liegt bei 0,7. Die angedeutete bimodale Response könnte ein Effekt sein, der durch Konkurrenz entsteht. tmin2 zeigt eindeutig, dass Eichen-Hainbuchen-Wälder bei Temperaturen um 5°C ihr Optimum haben, aber auch Frost bis -10°C unter Umständen vertragen. An der Responsekurve von prec6 kann abgelesen werden, dass durchschnittliche Niederschläge von 50 mm bis 100 mm im Juni das Optimum darstellen. Im Klimawandelszenario (Abbildung 4.10) zeigt das Modell eine Festigung des Habitats in Polen und eine klare Verschiebung der Eichen-Hainbuchen-Wälder nach Osten. Die Bedingungen in Frankreich verschlechtern sich und verbessern sich in Richtung Baltikum und Südschweden. Den Rasterzellen am Alpenrand und im Erzgebirge (montane Stufe) werden nun auch hohen Wahrscheinlichkeiten zugeordnet. Dies könnte einen Hinweis darauf geben, dass die klimatischen Bedingungen im Gebirge sich zugunsten der Hainbuche verbessern.

4.2.3 Französisch-Süddeutsche Eichen-Hainbuchen-Wälder



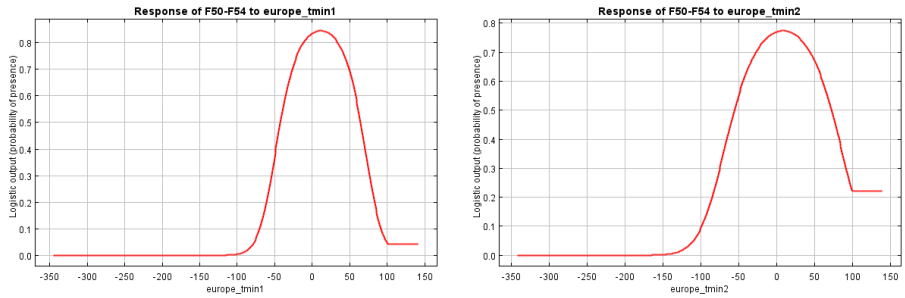
(a) Maximum-Entropie Habitatmodell der französich-süddeutschen Eichen-Hainbuchen-Wälder. Bestimmung des Kernhabitats durch Anwendung eines Schwellenwerts (0,5).



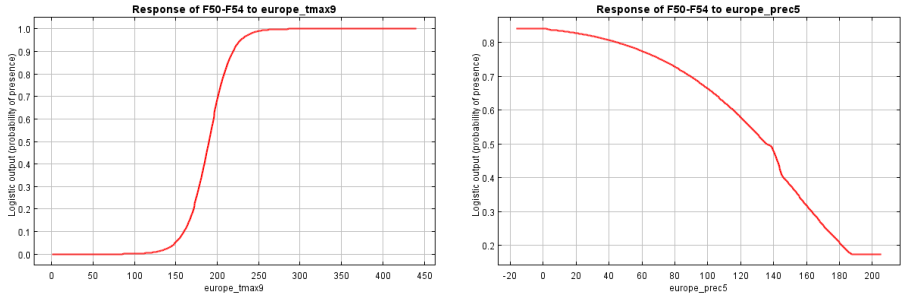
(b) Projektion des MaxEnt-Modells für die Vegetationseinheiten F50-F54 auf simulierte Klimadaten für 2080 (IPCC-Szenario A2a)

Abbildung 4.12: Modellvorhersage für die französisch-süddeutschen Eichen-Hainbuchen-Wälder (F50-F54).

Betrachtet man nur die westlichen der mitteleuropäischen Eichen-Hainbuchenwälder (F50-F54), so stellt sich die Situation ähnlich zu F50-F59 dar. Die in diesem Teilmodell gelernte Verteilung deckt sich für diesen Bereich mit dem Habitat des F50-F59-Modell.
Für dieses Modell hat MAXENT allerdings andere Variablen mit einem größeren prozentualen Beitrag zum Modell berechnet. An erster und zweiter Stelle stehen tmin1 und tmin2. Darauf folgen tmax9 und prec5. Der unimodale Responseverlauf



(a) tmin1, 23,6% Modellbeitrag (b) tmin2, 15,6% Modellbeitrag



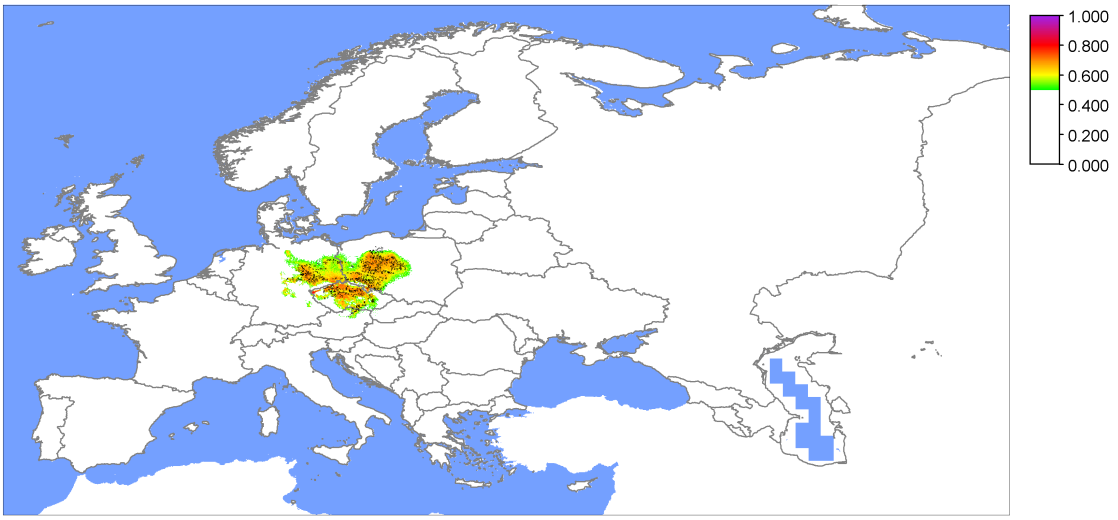
(c) tmax9, 8,3% Modellbeitrag (d) prec5, 8,1% Modellbeitrag

Abbildung 4.13: Responsekurven der wichtigsten Variablen von F50-F54. Temperaturvariablen in °C·10. Niederschlag in mm.

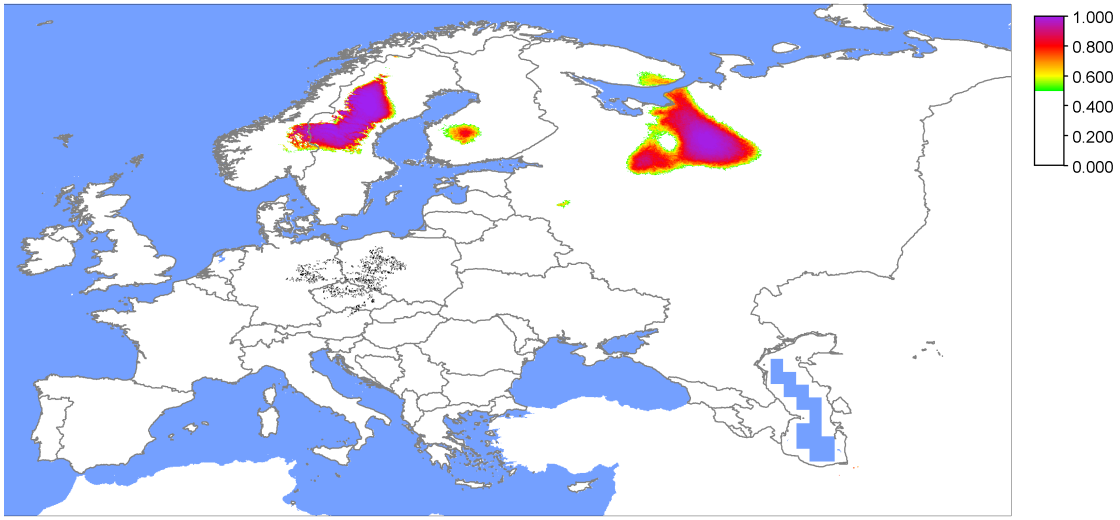
von tmin1 und tmin2 legt nahe, dass sowohl zu tiefe als auch zu hohe Temperaturen im Winter die Ausbildung der Gesellschaft nicht begünstigen. Die Optimaltemperatur liegt um 0°C.

Die Maximaltemperatur im September zeigt einen logistischen Verlauf. Unterhalb von circa 20°C sind die Bedingungen für Eichen-Hainbuchen-Wälder schlecht. Oberhalb dieser Schwelle zeigen die Wälder ein breites Spektrum und treten auch bei höheren Septembertemperaturen auf. Der Verlauf von prec5, dem mittleren Niederschlag im Mai, spiegelt die Wasserhaushaltspanne der Eichen-Hainbuchen-Wälder wieder. Im Lehrbuch von Fischer [16] wird dieser Zusammenhang für Galio-Carpinetum im nordbayrischen Keuperbergland charakterisiert. Dort heißt es : „Sind zudem die Niederschlagsmengen gering ([...]), so können sich Hainbuche und Winter-Linde gegenüber der Rotbuche durchsetzen.“ Diese Verhalten spiegelt sich auch im Responseverlauf für den Niederschlag im Juni wieder. Die Vorkommenswahrscheinlichkeit verringert sich erst bei größeren Niederschlägen. Diese Beobachtung deckt sich auch mit anderen Literaturangaben, in der die Wasserhaushaltsspanne mit „trocken bis mäßig trocken“ angegeben wird (vgl. [6]). Das auf 2080 projezierte Modell zeigt eine Konsolidierung des Areals, vor allem in Süddeutschland. Dort verbessert sich die Habitatgüte sehr stark. Das potentielle Habitat erweitert sich in Deutschland bis an den Südrand der Mittelgebirge, erstreckt sich über Nord- und Ost-Österreich und fast ganz Tschechien. Nur in Frankreich ist ein Rückgang des Habitat abzulesen.

4.2.4 Hercynisch-Polonische Eichen-Hainbuchen-Wälder



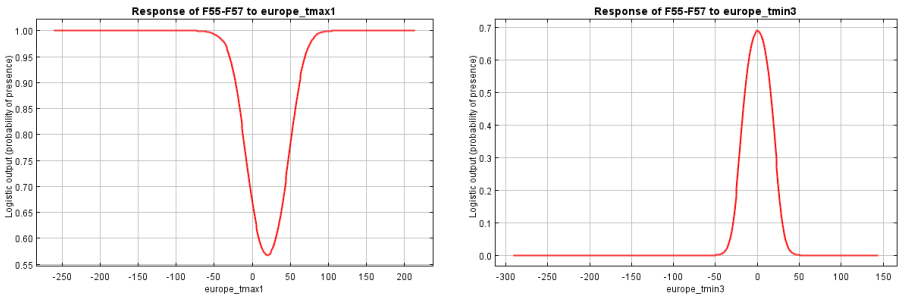
(a) Maximum-Entropie Habitatmodell der hercynisch-polonischen Eichen-Hainbuchen-Wälder. Bestimmung des Kernhabitats durch Anwendung eines Schwellenwerts (0,5).



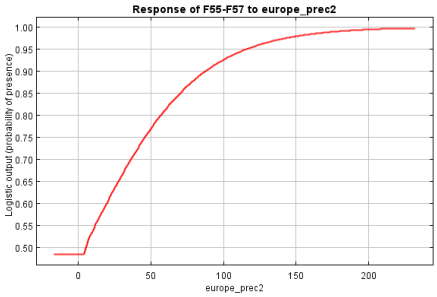
(b) Projektion des MaxEnt-Modells für die Vegetationseinheiten F55-F57 auf simulierte Klimadaten für 2080 (IPCC-Szenario A2a)

Abbildung 4.14: Modellvorhersage für die hercynisch-polonischen Eichen-Hainbuchen-Wälder (F55-F57).

Das Modell der Gesellschaften F55-F57 ($AUC = 0,996, R2 = 0,49$) zeigt ein sehr kompaktes Habitat, das sich in Ostdeutschland, Tschechien und Westpolen befindet. MAXENT zeigt eine starke Abhängigkeit von tmax1, tmin3 und prec2. Der Responseverlauf von tmax1 ist umgekehrt unimodal mit einem Mittelwert von circa 0°C. Eichen-Hainbuchen-Wälder in dieser Region werden, nach dem Modell, nur durch Temperaturen um den Gefrierpunkt in ihrer Verbreitung eingeschränkt. Die minimale März-Temperatur (tmin3) zeigt eine unimodale Response mit einem Mittelwert



(a) tmax1, 39,7% Modellbeitrag (b) tmin3, 21% Modellbeitrag

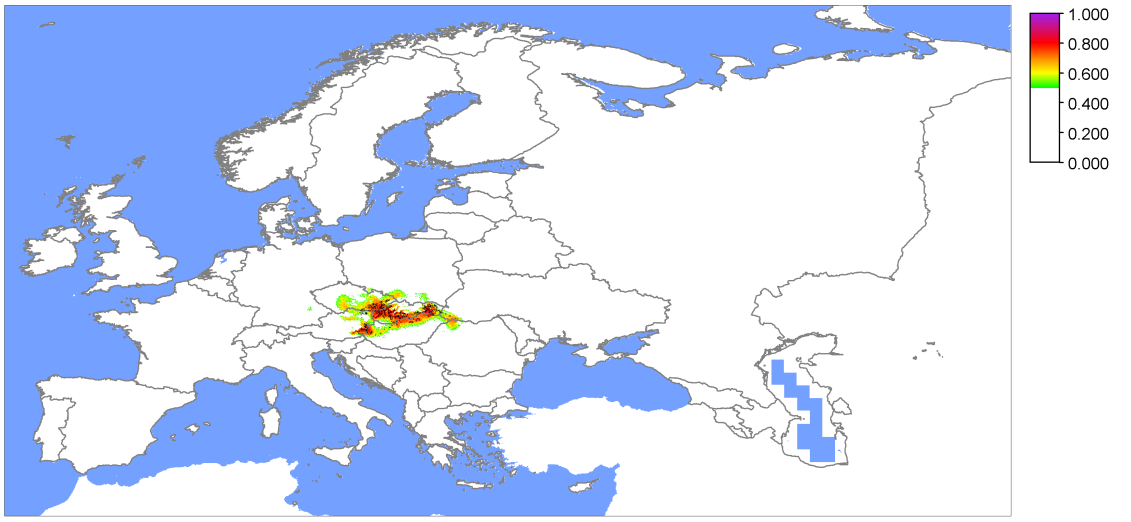


(c) prec2, 4,6% Modellbeitrag

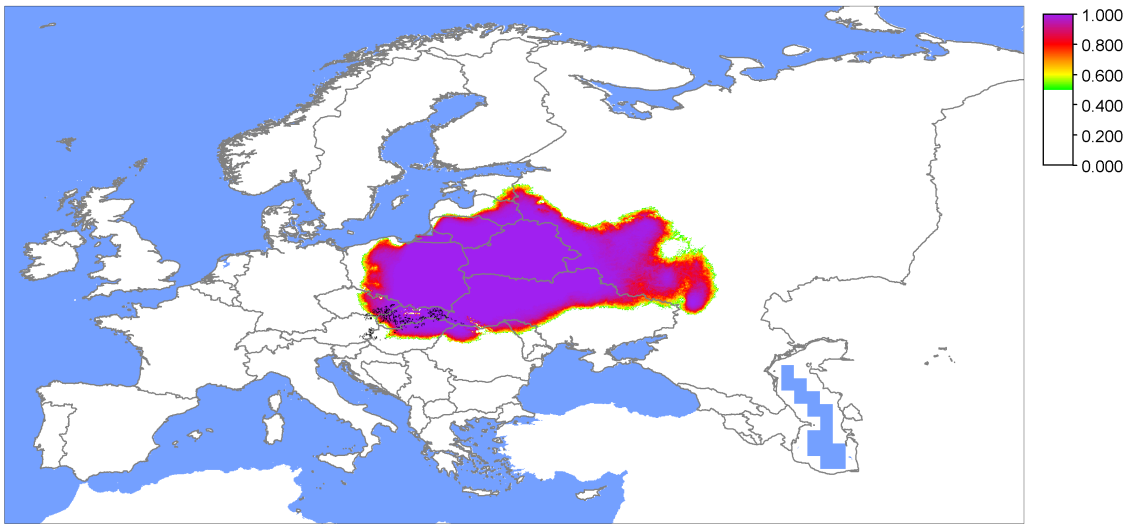
Abbildung 4.15: Responsekurven der wichtigsten Variablen von F55-F57. Temperaturvariablen in °C·10. Niederschlag in mm.

bei ungefähr 0°C. In Regionen, in denen die mittlere März-Temperatur nicht um den Gefrierpunkt liegt, treten seltener Eichen-Hainbuchen-Wälder auf. prec2 zeigt eine monoton ansteigende Response, je mehr Regen fällt, desto besser ist dies für die Wälder. Hier scheinen Eichen-Hainbuchen-Wälder auf Niederschlag anders zu reagieren (siehe Abschnitt 4.2.3). Für das Stellario-Carpinetum wird von Fischer [16] als Charakterisierung angegeben (Box Seite 263): „Wesentlicher Standortfaktor des Stellario-Carpinetum im (sub-)ozenaischen Teil Mitteleuropas ist die Bodenvernässung, durch welche die Rot-Buche stark zurückgedrängt wird.“ Das Modell deckt sich auch hier mit dem bekannten Wissen. Für das Jahr 2080 zeigt das Modell ein unerwartetes Verhalten. Es treten zwei potentielle Habitate nördlich des 60. Breitengrades auf, in einer Region, die natürlicherweise Tundren oder Nadelwälder als Klimaxgesellschaft besitzen. In Mitteleuropa scheinen sich die Bedingungen für die Gesellschaften stark zu verschlechtern, da hier keine großen Wahrscheinlichkeiten mehr auftreten.

4.2.5 Slovakische Eichen-Hainbuchen-Wälder



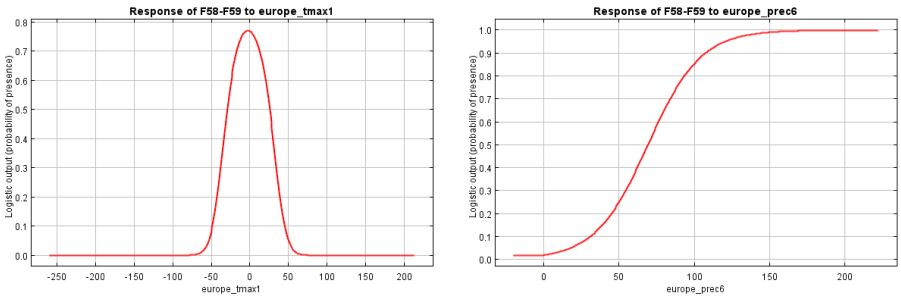
(a) Maximum-Entropie Habitatmodell der slovakischen Eichen-Hainbuchen-Wälder. Bestimmung des Kernhabitats durch Anwendung eines Schwellenwerts (0,5).



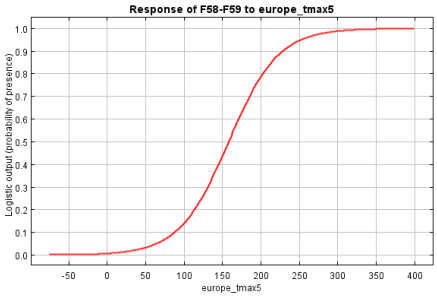
(b) Projektion des MaxEnt-Modells für die Vegetationseinheiten F58-F59 auf simulierte Klimadaten für 2080 (IPCC-Szenario A2a)

Abbildung 4.16: Modellvorhersage für die slovakischen Eichen-Hainbuchen-Wälder (F58-F59).

Das modellierte Habitat der Vegetationseinheiten F58-F59 ($AUC = 0,998, R2 = 0,67$) umfasst fast die ganze Slovakei, bis auf den Norden. Ausläufer erstrecken sich nach Nordungarn und in den Westen Tschechiens.
Die wichtigsten Variablen dieses Modells sind tmax1, prec6 und tmax5. Dabei zeigt tmax1 deutlich eine starke unimodale Abhängigkeit der Vorkommenswahrscheinlichkeit von der Maximaltemperatur im Januar an. Zu kalte, aber auch zu warme Januartemperaturen haben einen negativen Einfluss auf das Areal. Die anderen bei-



(a) tmax1, 36,6% Modellbeitrag (b) prec6, 9,9% Modellbeitrag



(c) tmax5, 9,9% Modellbeitrag

Abbildung 4.17: Responsekurven der wichtigsten Variablen von F58-F59. Temperaturvariablen in °C·10, Niederschlag in mm.

den wichtigen Faktoren weisen einen logistischen Verlauf auf, welcher nahe legt, dass Eichen-Hainbuchen-Wälder im Sommer eine Mindestmenge (mind. 50-60mm) an Niederschlag benötigen, außerdem Temperaturen von mindestens 10-15°C.

Das auf Umweltdaten von 2080 projizierte Habitat ist im Vergleich zum gelernten Modell sehr viel größer. Es umfasst alle Gebiete zwischen Polen und Westrussland. Im Norden dehnt es sich bis Estland aus. Im Süden erstreckt sich das Habitat bis in den Norden Rumäniens und umfasst den Nordteil der Ukraine. Die klimatischen Bedingungen scheinen sich für diese Gesellschaften stark zu verbessern.

4.3 Diskussion

Alle Modelle für europäische Eichen-Hainbuchen-Wälder zeigen im Vergleich zu den Läufen aus Abschnitt 4.1 sehr große Werte für R^2 und eine gute Diskriminanz (siehe Tabelle 4.8).
Die Vorhersagen des aktuellen potentiellen Habitats sind alle konsistent und unterscheiden sich nur an den Rändern. Es ist zu beobachten, dass die beiden Modelle für F34-F69 und F50-F59 stärker generalisiert haben, da sie im Vergleich zu den Teilmodellen eine größere Anzahl von nicht in der Stichprobe enthaltenen Flächen mit

Modell	N	R^2	**AUC**	wichtige Variablen
F34-F69	5615	0,39	0,952	tmax1, tmin12, prec6
F50-F59	2220	0,51	0,984	tmin2, tmax1, prec6
F50-F54	856	0,58	0,995	tmin1, tmin2, tmax9, prec5
F55-F57	916	0,49	0,996	tmax1, tmin3, prec2
F58-F59	448	0,67	0,998	tmax1, prec5, tmax5

Tabelle 4.8: Übersicht über die Güten der Modelle aus Abschnitt 4.2

hohen Wahrscheinlichkeiten belegen. In den Teilmodellen wird kaum generalisiert. Das vorhergesagte Habitat liegt sehr eng an den Stichprobenpunkten. Diese Beobachtung korreliert mit dem für MaxEnt-Modelle überdurchschnittlich großen Werten für R^2. Außerdem lässt sich beobachten, dass die Nischenbreiten der Teilmodelle sehr klein sind. Diese beiden Beobachtung könnten einen Hinweis darauf geben, dass die Teilmodelle zu stark angepasst sind (overfitting). Die Projektionen dieser Modelle auf das Jahr 2080 sollten deshalb kritisch betrachtet werden.

Die von MaxEnt gelernten Modell spiegeln in ihrer Variablenauswahl das bekannte Wissen über Eichen-Hainbuchen-Wälder gut wieder. In jedem Modell hat mindestens eine Wintertemperaturvariable einen großen Einfluss auf das Areal. Das Verhalten der Gesellschaft in Bezug auf Niederschlag ist meistens zweitrangig für die Verteilung. Die Response auf die Niederschlagsvariablen zeigt zum Beispiel für prec6 (F50-F59-Modell), dass Eichen-Hainbuchen-Wälder ihren optimalen Bereich bei relativ wenig Niederschlag haben, und bei größeren Niederschlägen häufig eine geringere Response aufweisen.

In allen Modellen, außer F55-F57, ist eine Sommerniederschlagsvariable wichtig. Dies deckt sich auch mit den schon bekannten Informationen (vgl. Seite 267 in [7]), dass größere Niederschläge im Sommer für Eichen-Hainbuchen-Wälder charakteristisch sind.

Alle Projektionen zeigen einen Osttrend. Die Vorhersagen sind zwischen F50-F59 und den Teilmodellen allerdings nicht konsistent, in der Projektion für F50-F59 spiegeln sich die vorhergesagten Gebiete aus den Teilmodellen nur zum Teil wieder. Dies lässt sich durch die unterschiedliche Anpassung der Modelle erklären. Für die westeuropäischen Eichen-Hainbuchen-Wälder hat MaxEnt andere Variablen für wichtig erachtet als zum Beispiel für die slovakischen Gesellschaften. Daraus kann geschlossen werden, dass die Teilgesellschaften auf Klimaveränderung unterschiedlich reagieren.

Bei der Interpretation der Habitatverschiebung sollte beachtet werden, dass Gesellschaften und keine einzelne Arten modelliert wurden. Da diese aus einer Menge von Arten bestehen, ist es zum Beispiel durchaus plausibel, dass die Gesellschaften

F50-F59 in Frankreich im Vergleich zum Jetzt-Modell kleinere Wahrscheinlichkeiten zugeordnet bekommen. Dies bedeutet nicht, dass alle an der Gesellschaft beteiligten Arten schlechtere Bedingungen vorfinden, sondern lediglich, dass die Gesellschaft in dieser Ausprägung dort wahrscheinlich nicht mehr zu finden sein wird, weil das Habitat einzelner Arten sich verändert. Umgekehrt zeigen hohe Wahrscheinlichkeiten zwar gute klimatische Bedingungen für Eichen-Hainbuchen-Wälder. Deren Ausbreitung in diesen Gebieten hängt aber auch von vielen weiteren Faktoren ab. Daher kann das Modell für 2080 nur eine mögliche Annäherung an die tatsächliche Veränderung des Habitats darstellen.

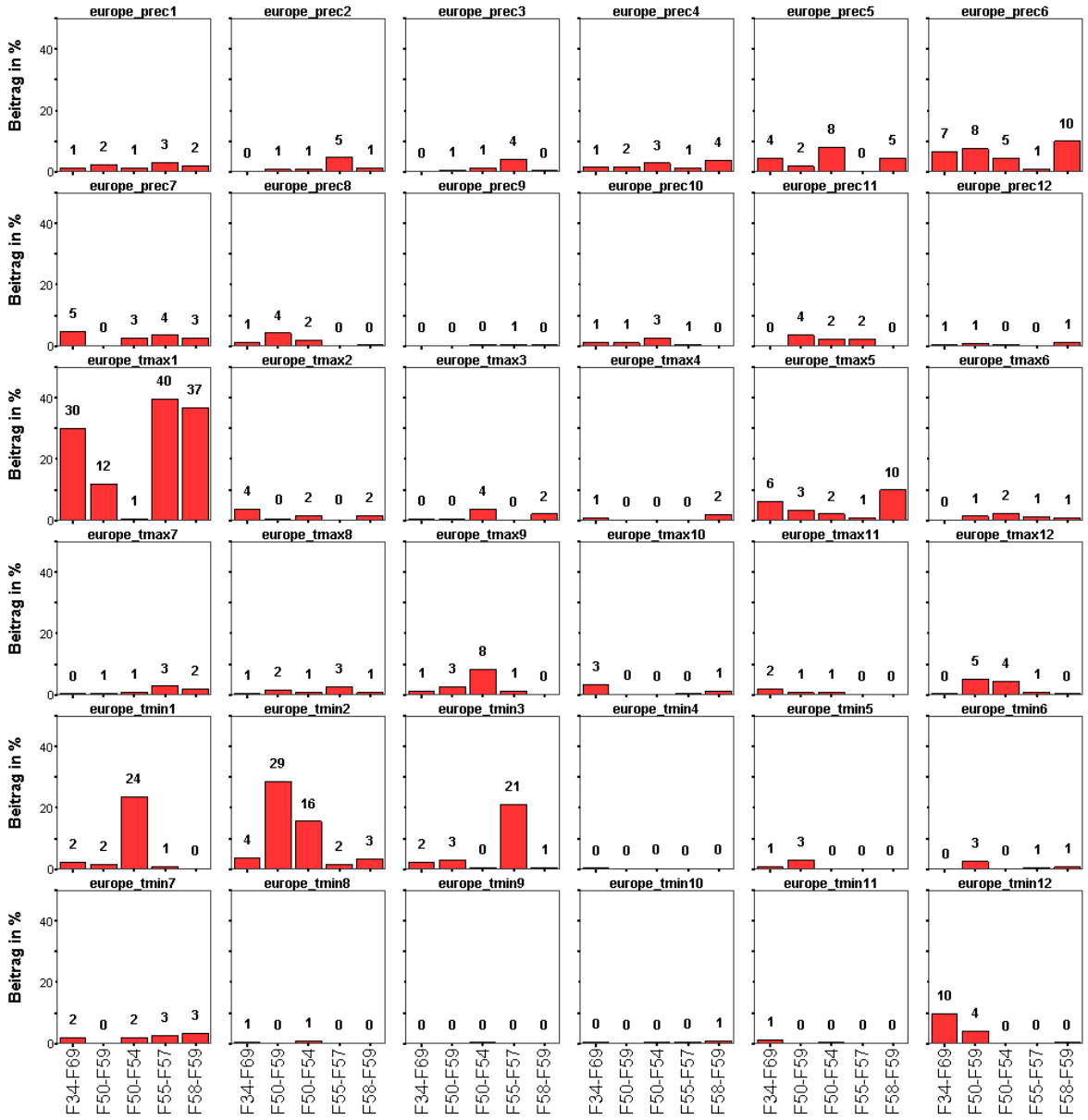


Abbildung 4.18: Prozentualer Beitrag aller 36 Variablen an den Modellen aus Abschnitt 4.2. Die Zahlen über den Balken geben den prozentualen Beitrag (auf ganze Zahlen gerundet) am Modell an.

5 Anwendung des Genetischen Programmierens

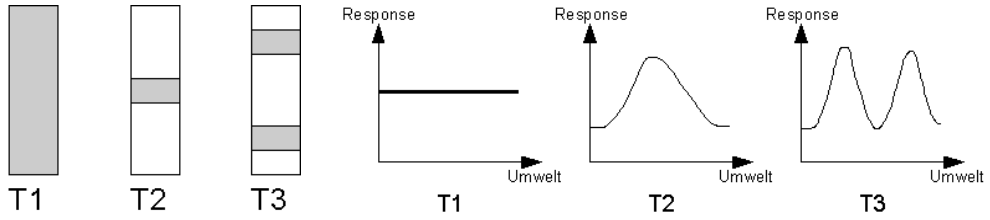
Bevor GP auf die Eichen-Hainbuchendaten angewandt wird, soll evaluiert werden, wie das HabitatGP Lösungen für Testprobleme findet.

5.1 Künstliche Testdaten

Es wurden drei Testprobleme konstruiert. Diese basieren alle auf einer künstlichen Umweltvariablen. Diese ist in einer 200×20-Rasterkarte mit Zellgröße 1 gespeichert. In x-Richtung sind alle Zellen gleich belegt, in y-Richtung verringert sich der Wert in jeder Zeile um 1. Die erste Zeile ist mit dem Wert 200 belegt (siehe Abbildung 5.1).

```
nrows 200
ncols 20
xllcenter 0.0
yllcenter 0.0
cellsize 1.0
nodata_value -9999
200 200 200 200 200 200 200 200 200 200 200 ...
199 199 199 199 199 199 199 199 199 199 199 ...
198 198 198 198 198 198 198 198 198 198 198 ...
............
3 3 3 3 3 3 3 3 3 3 3
2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1
```

Abbildung 5.1: Künstliche Umweltvariable: Auszug aus der Datei artificialmap.asc



(a) Sampling (b) Erwartete Response

Abbildung 5.2: Schematische Darstellung der Testsamplings. Die graue Fläche zeigt den Bereich aus dem alle Rasterpunkt als Positivstichprobe genommen wurden.

Durch das Sampling einer virtuellen Art wird bestimmt, ob die Art in Bezug auf diese künstliche Umweltvariable eine konstante (T1), unimodale (T2) oder bimodale (T3) Response besitzt. Die konstante Response wird erzeugt, indem alle Punkte der Karte in der Lernstichprobe enthalten sind. Ein Sampling, dass nur Punkte aus einem zusammhängenden Bereich (z.B. der Mitte) enthält, hat eine unimodale Response zur Folge. Die bimodale Response wird erreicht, indem zwei verschiedene Bereiche der Karte als Stichprobe dienen. Die Samplings sind in Abbildung 5.2 schematisch dargestellt. Da die Werte in der künstlichen Karte räumlich sortiert sind, entspricht die Vorhersage der Wahrscheinlichkeiten für die Karte direkt der Responskurve. Umgekehrt geben die dargestellten Responsekurven die Wahrscheinlichkeitsverteilung über die Karte wieder.

Für die Anwendung auf die Testdaten wurde das GP mit drei verschiedenen Zielfunktionen konstruiert. Diese setzten sich aus der Bewertung der Anzahl der Hits, der Anzahl der verwendeten Variablen, der Abweichung von den Erwartungswerten und der Differenz zur maximalen Entropie zusammen. Folgende Zielfunktionen werden betrachtet: Hits & Variablenanzahl (FH), Entropie und Erwartungswerte (EntExp), Variablenanzahl, Hits, Entropie und Erwartungswerte (Full).

$$Z_{FH}(ind) = \frac{|ps| - hits}{|ps|} + \frac{\#Variablen - \#verwendeteVariablen}{\#Variablen} \tag{5.1}$$

$$Z_{EntExp}(ind) = \frac{H_{max} - H_{ind}}{H_{max}} + \sum_f \frac{\widehat{\pi}[f] - \widetilde{\pi}[f]}{\widehat{\pi}[f]} + \sum_f \frac{\widehat{\pi}[f]^2 - \widetilde{\pi}[f]^2}{\widehat{\pi}[f]^2} \tag{5.2}$$

$$Z_{Full}(ind) = Z_{FH}(ind) + Z_{EntExp}(ind) \tag{5.3}$$

Um die Qualität der verschiedenen Zielfunktionen auszuwerten, wurden für 10 Replikate (Läufe mit gleichem Parametersatz) die Laufzeit und die erreichte Fitness (Optimum 0) ermittelt und die Übereinstimmungen der besten Individuen mit der erwarteten Response gezählt. Die erwartete Response wird qualitativ bewertet, da den Testproblemen keine Funktion zu Grunde liegt, welche reproduziert werden soll, sondern nur ein bestimmte Form der Response erwartet wird. Die Tabellen 5.2,

5.3 und 5.4 geben die gemittelten Fitnesswerte und Laufzeiten der verschiedenen Zielfunktionen wieder. Für die Berechnung der Mittelwerte wurde jeweils die Fitness des besten gefundenen Individuums jedes Replikats verwendet. Die MAXENT-Parameter für diese Läufe sind in Tabelle 5.1 nachzulesen.

Parameter	**Wert**
Populationsgröße	100
Generationen	100
Anzahl eingefügter Mutationen	10
Anzahl Turniere pro Generation	5
Mutationswahrscheinlichkeit	0,6
Rekombinationswahrscheinlichkeit	1
Anzahl Rechenregister	10
Anzahl Konstanten	10
Anzahl alle Register	21
Hit-Schwelle	0,9
Anzahl Replikate	10

Tabelle 5.1: Parametersatz für MAXENT für die Läufe in Abschnitt 5.1

Lernverlauf

Die Grafiken zu der Auswertung des Lernverlaufs befinden sich im Anhang. Für jedes Testproblem und Zielfunktion wurde die mittlere Entwicklung der Population im Bezug auf die Komponenten der Zielfunktion berechnet und dargestellt. Dazu wurden jeweils zehn Generationen und die entsprechenden Populationen aus den Replikaten zusammengefasst und gemittelt.
Die Entwicklung der Hits über die Generationen zeigt, dass in den Läufen von Z_{FH} und Z_{Full}, welche die Anzahl der Hits in die Fitness mit einbeziehen, die mittlere Anzahl der Hits stetig ansteigt. Für Z_{Full} lässt sich das sehr deutlich erkennen. Allerdings ist das Konfidenzintervall für alle Werte relativ groß. Dies spricht für eine große Diversität in der Population. Eine Eigenschaft, die dieses Verhalten begünstigt, ist, dass die Gleichverteilung maximale Anzahl von Hits erzeugt.
Alle Versuche zeigen eine Verbesserung der mittleren Fitness im Verlauf einer Optimierung. Die Fehlerbalken sind für Z_{FH} und Z_{EntExp} meist größer als die Fehlerbalken für Z_{Full}. Das spricht dafür, dass die Läufe mit Z_{Full} in den vorgegebenen 100 Generationen noch nicht ihr Optimum erreicht haben und für Z_{Full} mehr Generationen Individuen mit einer besseren Fitness erzeugen könnten.

Die Entropie zeigt für alle Läufe einen konstanten Verlauf. Die mittlere Entropie liegt immer sehr nah an der maximalen Entropie ($\ln(|X|) = \ln(4000) = 8,29$). Die Fehlerbalken zeigen, dass über die verschiedenen Läufe kaum Abweichungen auftreten, ganz gleich ob die Entropie in die Zielfunktion eingeht oder nicht. Die hohe Entropie zeigt an, dass die meisten Modelle der Gleichverteilung (konstante Funktion) nahe kommen. Die gelernten Funktionen sind allerdings nicht konstant, sondern nutzen eine numerische Instabilität im Programm aus.
Je nachdem, ob die Abweichung von den empirischen Mittelwerten bzw. Varianzen in die Zielfunktion eingehen, lässt sich klar beobachten, dass die Abweichungen für Z_{FH} stagnieren oder sogar ansteigen. Die beiden Zielfunktionen Z_{EntExp} und Z_{Full}, die die Abweichungen mitbewerten zeigen eine eindeutige Verbesserung im Lernverlauf (die Abweichung wird kleiner).

Auswertung der gelernten Funktion

Da alle Variablen vor der Optimierung auf das Intervall [0,1] normiert werden, wurde für diese Auswertung der Funktionsverlauf nur in diesem Intervall bewertet. Das Verhalten der Modellfunktion außerhalb dieses Wertebereichs hat auf das Modell keinen direkten Einfluss.
Die in T1 erwartete konstante Funktion wurde von den verschiedenen Zielfunktionen relativ häufig gefunden (siehe Tabelle 5.2). Dabei hat Z_{FH} die beste Trefferquote. Die gefundenen Funktionen sind meist sehr kompliziert und erzeugen vor der Normierung sehr große absolute Werte ,wie zum Beispiel:

$$9 \cdot 10^{-\frac{e^7}{3}} \cdot e^{-\frac{7e^7}{6}} \approx 1 \times 10^{191}.$$

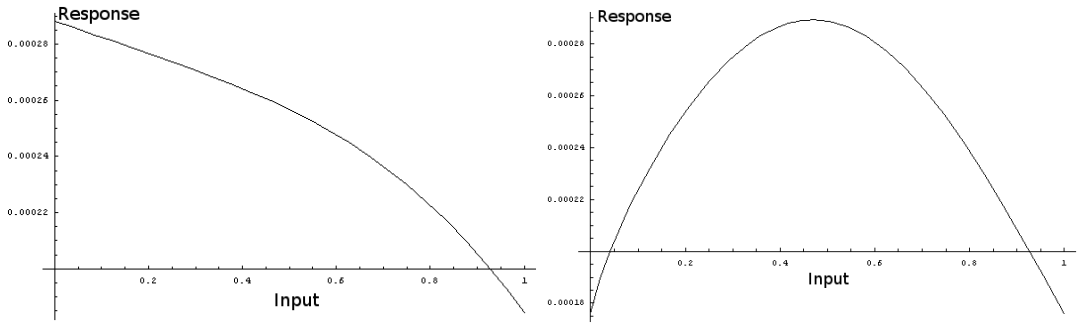
Für T1 ist Z_{FH} eigentlich nicht die optimale Zielfunktion, da sie die Anzahl der vom Genom verwendeten Variablen mit berücksichtigt. Konstante Funktionen können aber nur entstehen, wenn die Variable nicht vorkommt oder durch ein semantisches Intron nicht in die Berechnung eingeht (z.B. Multipikation mit 0).

Zielfunktion	mittlere Fitness	Laufzeit	erwartete Response
FH	0,008 (0,007)	1008s[1] (218,7s)	6/10
EntExp	0,038 (0,007)	1909 (14,9s)	2/10
Full	0,1782 (0,068)	2369s (120,9s)	3/10

Tabelle 5.2: Ergebnisse (beste Individuen) der verschiedenen Zielfunktionen aus der Anwendung auf T1. Alle Werte gemittelt aus den Werten der 10 besten Individuen. Die Werte in Klammern sind die Standardfehler der Mittelwerte.

Für das Testproblem T2 (siehe Tabelle 5.3) lässt sich beobachten, dass Z_{FH} die gesuchte Funktion nicht lernen konnte. Mit Z_{EntExp} und Z_{Full} hat das HabitatGP ebenfalls kaum gute Resultate erzielt. Eine der Funktionen, die von Z_{EntExp} gelernt wurden und annähernd einen unimodalen Verlauf zeigten entspricht nicht exakt der erwarteten Response, zeigt aber fast den gewünschten Verlauf (siehe Abbildung 5.3(a)). Das beste Individuum, das von Z_{Full} gefunden wurde, entspricht genau der erwarteten Response (siehe Abbildung 5.3(b)) und hat die funktionale Form:

$$\frac{1 + 16 \cdot (x^{x+1})^{-x}}{96674.1}.$$



(a) Z_{EntExp}, Fitness = 0,146329, Entropie = 8,28742, Hits = 0

(b) Z_{Full}, Fitness = 0,226455, Entropie = 8,28445, Hits = 779

Abbildung 5.3: Darstellung von zwei Individuen als Responsekurve. Die y-Achse beschreibt die Rohwahrscheinlichkeit. Die x-Achse ist der Wert der Inputvariablen, normiert auf [0,1]. Hier entspricht die x-Achse der künstlichen Umweltkarte artificialmap.asc und der Verlauf der Response der geographischen Verteilung

Zielfunktion	mittlere Fitness	Laufzeit	erwartete Response
FH	0,095 (0,095)	134,7s (44,7s)	0/10
EntExp	0,1639 (0,007)	549s (27,97s)	2/10
Full	0,2047 (0,005)	618s (20,19s)	1/10

Tabelle 5.3: Ergebnisse der verschiedenen Zielfunktionen aus der Anwendung auf T2. Alle Werte gemittelt aus den Werten der 10 besten Individuen. Die Werte in Klammern sind die Standardfehler der Mittelwerte.

Für T3 hat nur Z_{EntExp} in einem von 10 Läufen eine Lösung gefunden, die die erwartete Lösung annähert.

Zielfunktion	mittlere Fitness	Laufzeit	erwartete Response
FH	0,1583 (0,108)	339s (81,95s)	0/10
EntExp	0,1215 (0,005)	486 (19,02s)	1/10
Full	0,2210 (0,052)	505s (19,47s)	0/10

Tabelle 5.4: Ergebnisse der verschiedenen Zielfunktionen aus der Anwendung auf T3. Alle Werte gemittelt aus den Werten der 10 besten Individuen. Die Werte in Klammern sind die Standardfehler der Mittelwerte.

Die gelernte Funktion ist

$$\frac{(x^2)^{x^2}}{3230.68}.$$

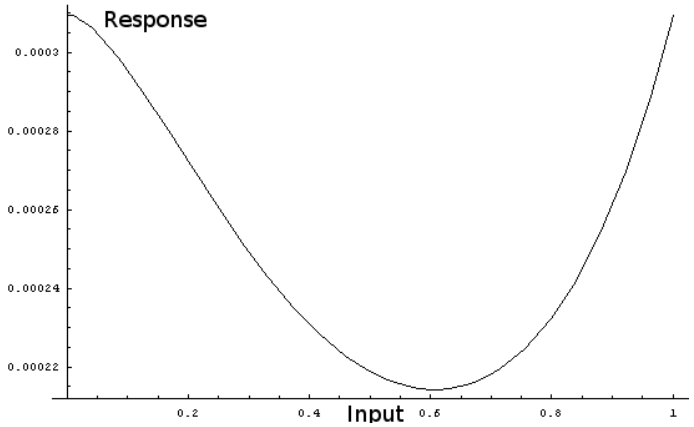


Abbildung 5.4: Phänotyp der besten Individuum für T3 mit Z_{Full}, Fitness = 0,117653, Entropie = 8,28645, Hits = 42. Die y-Achse beschreibt die Rohwahrscheinlichkeit. Die x-Achse ist der Wert der Inputvariablen, normiert auf [0,1]. Hier entspricht die x-Achse der künstlichen Umweltkarte artificialmap.asc und der Verlauf der Response der geographischen Verteilung.

Sehr häufig konnte beobachtet werden, dass das HabitatGP eine einfachere Funktion (z.B. lineare) gelernt hat, die das Sampling nur zum Teil gut wiedergibt. Eine andere häufig gemachte Beobachtung ist, dass die funktionale Form zum Beispiel auf quadratische oder kompliziertere Responseverläufe schließen lässt, sich im ausgewerteten Intervall aber dann häufig linear oder sogar fast konstant verhält. Die interessanten Eigenschaften dieser Funktionen liegen meist außerhalb des Intervalls [0,1].

5.2 Eichen-Hainbuchen-Daten

Um dem erhöhten Rechenaufwand bei großen Karten entgegen zu wirken, wurde das HabitatGP zunächst auf einen verkleinerten Kartenausschnitt angewandt, welcher nur Deutschland und Teile der Nachbarländer enthält. Als Stichprobe dienten die Präsenzpunkte der Gesellschaften F50-F59. Stichprobenpunkte, die außerhalb dieses verkleinerten Untesuchungsgebietes liegen, wurden automatisch durch das HabitatGP nicht in die Optimierung einbezogen. Die Menge der Umweltvariablen wurde auf bio1, bio2 und bio3 beschränkt.
Ein Lauf mit Z_{FH} hat zum Beispiel eine einfache lineare Modellfunktion $M(\mathbf{f}(x)) = bio1(x) * 6$ erzeugt. Die zugehörige Verteilung ist in Abbildung 5.5 zu sehen. Dabei wurde nur eine der drei Variablen überhaupt verwendet und die daraus resultierende Verteilung lässt keine Aussage über die Verteilung der Eichen-Hainbuchen-Wälder zu. Im Modell ist nur die Information enthalten, dass Eichen-Hainbuchen-Wälder nicht im Gebirge vorkommen. Vermittelt wird diese durch bio1, die Jahresmitteltemperatur.

5.3 Diskussion

Die unbefridigenden Ergebnisse des HabitatGP können mehrere Ursachen haben. Vorstellbar ist, dass durch einen konzeptionellen Fehler im Algorithmus bzw. der Programmierung die gewünschte Eigenschaft, gegen gute Lösungen zu konvergieren, ausgehebelt worden ist. Die Verlaufsdiagramme zeigen für alle Läufe allerdings eine Verbesserung aller Komponenten einer Fitnessfunktion über die Zeit. Daher ist es wahrscheinlicher, dass die gewählten Parameter bzw. konzeptionelle Fehler die Konvergenz zwar nicht komplett verhindern, aber verlangsamen und die Qualität der gefundenen Lösungen mindern.

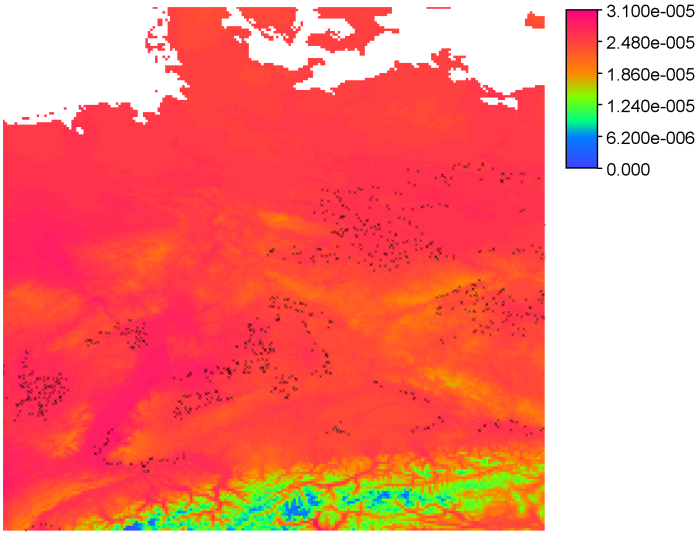


Abbildung 5.5: Ergebnis des HabitatGP auf einer Deutschlandkarte. Die schwarzen Punkte stellen die Stichprobe für die Gesellschaften F50-F59 in diesem Gebiet dar. Die Wahrscheinlichkeitsverteilung resultiert aus einer linearen Modellfunktion $M(\mathbf{f}(x)) = bio1(x) * 6$.

Der Lernerfolg der Testprobleme legt nahe, dass die vorgestellten Zielfunktionen, trotz ihrer Plausibilität, für diese Problemstellung nicht geeignet sind. Allerdings könnte dieser Eindruck auch entstehen, da der Suchraum sehr groß und komplex ist und das GP gar nicht die Möglichkeit hat gute Lösungen zu finden.
Stattdessen werden numerische Grauzonen ausgenutzt. Durch die lineare Transformation der Wahrscheinlichkeiten auf das Intervall [0,1] zu Berechnung der Hits tendieren alle Funktionen zur Gleichverteilung, da diese die maximale Anzahl an Hits erreicht. Die Hinzunahme des Entropie-Kriteriums scheint hier ein Ungleichgewicht herzustellen. Für T1 wurde zum Beispiel nicht zu 100% die Gleichverteilung gelernt, obwohl dies die Verteilung mit der größten Entropie ist.
Der große Suchraum führt dazu, dass das GP zwar sehr komplexe Funktionen lernt, diese aber im ausgewerteten Bereich [0,1] meist ein sehr einfaches Verhalten zeigen (konstant oder linear). Bei sehr großen Karten kommt noch ein Speicherplatzproblem hinzu. Ein Modell für ganz Europa führt in der Speicherverwaltung durch das HabitatGP zu einem Überlauf, so dass das Programm abstürzt. Die Testläufe haben gezeigt, dass GP zum Lernen der Modellfunktion eines Habitatmodells anwendbar ist und funktioniert. Das Beispiel aus Kapitel 5.2 hat gezeigt, dass eine Anwendung auf echte Daten möglich ist, die Modellgüte aber noch stark verbesserungwürdig ist, so dass noch kein effektiver Einsatz zur Beantwortung echter Fragestellungen sinnvoll wäre. Hier besteht weiterer Forschungsbedarf.

6 Abschließende Diskussion und Ausblick

In dieser Arbeit wurden Habitatmodelle für europäischen Eichen-Hainbuchen-Wälder mit Hilfe der Maximum-Entropie-Methode (MaxEnt) erstellt. Die Maximum-Entropie-Methode wurde in verschiedenen Szenarien getestet und evaluiert. Die Evaluation basiert auf der Bewertung der Anpassung durch R^2, der Bewertung der Diskriminanz durch die Fläche unter der ROC-Kurve (AUC) und durch das kontinuierliche Boycemaß. Es hat sich gezeigt, dass MaxEnt auch aus kleinen Stichproben gute Modelle erzeugt (Kapitel 4.1.1. Für die Testläufe hat sich eine Stichprobengröße von 220 (10%) als optimal erwiesen. Der Einfluss von linear korrelierten Variablen ließ sich nur für negativ korrelierte Variablen nachweisen. Die Modelle, die auf negativ-korrelierten Variablen basieren hatten für alle Gütemaße eine signifikant besseren Wert. Die Signifikanz basiert auf einem nicht-parametrischen Test, da nicht für alle betrachtete Gruppen, die in die statistische Analyse eingegangen sind, eine Normalverteilung angenommen werden konnte. Nicht-parametrische Tests basieren in der Regel auf Rängen und sind etwas gröber als zum Beispiel t-Tests. Es könnte also sein, dass sich Modelle aus positiv korrelierten Variablen doch signifikant unterscheiden, wenn ein feinerer parametrischer Test angewandt wird. Um die Voraussetzungen dafür zu schaffen, könnte eine größere Datenbasis von Vorteil sein, zum Beispiel Modelle aus allen positiv korrelierten Variablen. Weiterhin könnten nichtlineare Korrelationen einen Einfluss haben. Diese Effekte könnten in den Modellen vorhanden sein, wurden durch die Untersuchung aber nicht erfasst. Das in diesem Szenario jeweils nur zwei Variablen in die Modelle eingegangen sind, hat sich für die Korrelationsuntersuchung in schlechten Modellgüten niedergeschlagen. Gehen mehr als zwei Variablen in ein Modell ein, können die vielfältigen Beziehungen zwischen den Variablen ganz verschiedene Einflüsse haben. Die Modellkonstruktion durch eine vorgeschaltete Diskriminanzanalyse hat ergeben, dass, im Vergleich zur Untersuchung der Stichprobengröße, nicht nur wenige Stichprobenpunkte, sondern auch wenige Variable ausreichen, um gute Modelle zu berechnen. Im Korrelationsszenario hat sich gezeigt, dass Modelle aus zwei Variablen noch nicht ausreichen,

um gute Modelle zu erstellen. Dieser Eindruck hat sich im Diskriminanzszenario bestätigt. Auch hier reichen zwei Variablen noch nicht aus, um gute Modelle zu erstellen. Bis zur sechsten Variable zeigt sich eine stetige Verbesserung, sowohl im AUC als auch in R^2. Ab der sechsten Variable zeigt sich keine signifikanten Besserung der Anpassung mehr. Der AUC-Wert kann hingegegen durch mehr Variable noch erhöht werden. Diese Erhöhung ist allerdings sehr gering, da schon das erste Modell einen sehr hohen AUC-Wert hat. Dass sich AUC stetig verbessert, die Anpassung aber nicht signifikant höher wird, könnte darauf hindeuten, dass die Modelle D7-D10 stärker zur Überanpassung neigen. Überanpassung tritt auf, wenn wenige Daten mit einem sehr komplexen Modell erfasst werden. Das Modell gibt dann diese wenigen Daten sehr genau wieder, hat aber kaum auf ähnliche Fälle generalisiert. Dies kann sich in der graphischen Darstellung der Verteilung zeigen, wenn die „Hauptmasse“ der Wahrscheinlichkeiten nur auf den Stichprobenpunkten liegt. Ein echter Nachweis von Überanpassung sind die hier gemachten Beobachtungen allerdings nicht, sondern eher ein Hinweis auf Überanpassung. Durch die einheitliche Generierung (gleiche Stichprobengröße, gleiche Variablenauswahl) der Modelle in Kapitel 4.1 konnten die vielen Daten auch verwendet werden um für MaxEnt eine Zusammenhang zwischen R^2 und AUC herzustellen. Es hat sich herausgestellt, dass sehr schlecht angepasste Modelle (die meisten Wahrscheinlichkeiten liegen um 0,5) ein breites Spektrum an AUC-Werten annehmen können. Die gute Diskriminanz scheint hier also unabhängig von der Anpassung zu sein. Sobald das Modell eine gewisse Anpassungsgüte erreicht hat, ist der AUC immer sehr hoch. Dieses Verhalten ließ sich am besten durch eine logistische Regression (Obergrenze bei 1) beschreiben.

Die Anwendung von MaxEnt auf verschiedene Eichen-Hainbuchen-Gesellschaften sollte mit der Ausrichtung der Anwendung auf Klimaszenariodaten für 2080 geschehen. Die Klimaszenariodaten sind bis heute nur für die Basisvariablen verfügbar, so dass die bio-Variablen nicht ohne größeren Aufwand für die Anwendung bereitgestellt werden konnten. Die Modelle spiegeln jeweils die Verteilung aber auch die Ansprüche der Eichen-Hainbuchen-Wälder an ihre Umwelt gut wieder, da die für das Modell wichtigen Variablen sich mit dem bekannten Wissen über Eichen-Hainbuchen-Wäldern decken. Die Projektion der größeren Modelle F34-F69 und F50-F59 auf das Jahr 2080 zeigen den Osttrend, der sich in allen Modellen wiederfindet. Die klimatischen Bedingungen scheinen sich in Westeuropa in einen für Eichen-Hainbuchen-Wälder ungünstigen Bereich zu verschieben und in Nord-Osteuropa zu verbessern. Während der Osttrend im Gesamtmodell (F34-F69) sehr deutlich ist und das potentielle Habitat sich nur noch auf Polen und das Baltikum beschränkt, zeigt das Modell der mitteleuropäischen Hainbuchen-Wälder eine Konsolidierung des Habitats in Polen und nur einen leichten Osttrend. Auf den ersten Blick sind

diese Vorhersagen nicht konsistent, was in der unterschiedlichen Anpassung an die Daten begründet liegt. So haben zwar beide Modelle tmax1 und prec6 als wichtigste Variablen gelernt, unterscheiden sich aber sowohl in der Wichtung (tmax1 ist einmal wichtigste und einmal zweiwichtigste Variable) als auch in der Auswahl der dritten Variable. Die unterschiedliche Wichtung führt dazu, dass die Modelle auf die Änderungen einer Variablen unterschiedlich reagieren.
Die MaxEnt-Methode hat sich als geeignetes Werkzeug zur Erstellung von Habitatmodellen bestätigt. Allerdings basiert die Darstellung der Responsekurven auf Schnitten durch die mehrdimensionale Responsefunktion, was dazu führt, dass die Response sich in Abhängkeit der Daten verändern kann, obwohl sie noch den selben Zusammenhang repräsentieren sollten. Eine Marginalisierung der Response (nicht in MAXENT implementiert) sollte auf jeden Fall ein stabileres Bild der Response liefern, indem die Verläufe über alle Parameterkombination gemittelt werden. Ein solches Verfahren ist für die hochdimensionalen MaxEnt-Modelle allerdings auch sehr rechenintensiv, wenn die Ergebnisse nicht zu grob sein sollen.
Die Verwendung von GP zur Berechnung von Habitatmodellen hat gezeigt, dass GP in der Lage ist einfache Testprobleme zu lösen, aber für den sinnvollen Einsatz noch nicht ausgereift genug ist. Dies liegt vor allem an der Größe des Suchraums. Dieser kann zwar durch eine Längenbeschränkung der Individuen verkleinert werden. Gute Lösungen, die oberhalb dieser Maximallänge liegen, werden dabei unter Umständen ausgeschlossen. Die Anwendung von erweiterten Konzepten, wie fitness sharing, implizite Längenbeschränkung (Länge als Term in die Zielfunktion einbeziehen), Schrittweitensteuerung und vielleicht Fitnessbewertung mittels Paretodominanz könnten Möglichkeiten sein, gute Lösungen effektiver zu finden.

Literaturverzeichnis

[1] *Definition Habitatmodell*. Lexikon Geoinformatik-Service, http://www.geoinformatik.uni-rostock.de/einzel.asp?ID=415207928. aktualisiert 22. August 2002, Besucht am: 30. Juli 2008.

[2] *GNU Scientific Library*. http://www.gnu.org/software/gsl/. Erstellt am: 18. Juli 2008; Besucht am: 30. Juli 2008.

[3] *EU-Richtlinie 92/43/EWG, Fauna-Flora-Habitatrichtline*, Mai 1992. Letzte konsolidierte Fassung 01. Januar 2007.

[4] *Desktop GARP*. http://www.nhm.ku.edu/desktopgarp/, März 2007. Besucht am: 20. Juli 2008.

[5] Austin, M. P.: *Spatial prediction of species distribution: an interface between ecological theory and statistical modelling*. Ecological Modelling, 157:101–118, 2002.

[6] Bayrische Landesanstalt für Wald und Forstwirtschaft: *LWF-Wissen / LWF-Bericht - Beiträge zur Hainbuche*, Bd. 12, 1996.

[7] Bohn, U., R. Neuhäusl, G. Gollub, C. Hettwer, Z. Neuhäuslova, T. Raus, H. Schlüter und H. Weber: *Karte der natürlichen Vegetation Europas. Maßstab 1:2500000*. Bundesamt für Naturschutz, Bonn, 2000/2003.

[8] Brameier, M.: *On Linear Genetic Programming*. Doktorarbeit, Universität Dortmund, 2004.

[9] Carnaval, A. C. und C. Moritz: *Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest*. Journal Of Biogeography, 35(7):1187–1201, 2008.

[10] Cherrett, J. M. (Hrsg.): *Ecological Concepts*. Blackwell Scientific Publications, 1989.

[11] Dormann, C. F., T. Blaschke, A. Lausch, B. Schröder und D. Sündgerath (Hrsg.): *Habitatmodelle - Methodik, Anwendung, Nutzen. Tagungsband zum Workshop vom 8.-10. Oktober 2003 am UFZ-Leipzig. UFZ-Berichte 9/2004*. UFZ - Umweltforschungszentrum Leipzig, Oktober 2004.

[12] Elith, J., C. Graham, R. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. Hijmans, F. Huettmann, J. Leathwick, A. Lehmann, J. Li, L. Lohmann, B. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. Overton, A. Peterson, S. Phillips, K. Richardson, R. Scachetti-Pereira, R. Schapire, J. Soberon, S. Williams, M. Wisz und N. Zimmermann: *Novel methods improve prediction of species' distributions from occurrence data*. Ecography, 29(2):129–151, 2006.

[13] Engleder, T.: *Ein Habitatmodell für den Luchs in der 3-Länder-Region Böhmerwald*. Diplomarbeit, Universität Wien, 2001.

[14] Ficetola, G. F., W. Thuiller und C. Miaud: *Prediction and validation of the potential global distribution of a problematic alien invasive species - the American bullfrog*. Diversity And Distributions, 13(4):476–485, 2007.

[15] Fielding, A. und J. Bell: *A review of methods for the assessment of prediction errors in conservation presence/absence models*. Environmental Conservation, 24(1):38–49, 1997.

[16] Fischer, A.: *Forstliche Vegetationskunde - Eine Einführung in die Geobotanik*. Parey Buchverlag, Berlin, 2002. 2. neubearbeitete und erweiterte Auflage.

[17] Graham, C. H. und R. J. Hijmans: *A comparison of methods for mapping species ranges and species richness*. Global Ecology And Biogeography, 15(6):578–587, 2006.

[18] Griesche, C.: *Die Eiche*. In: *SDW- Faltblätter*, Bd. 6. Schutzgemeinschaft Deutscher Wald Bundesverband e.V., 2007.

[19] Griesche, C. und S. Krömer-Butz: *Die Hainbuche*. In: *SDW- Faltblätter*, Bd. 10. Schutzgemeinschaft Deutscher Wald Bundesverband e.V., 2007.

[20] Guisan, A. und W. Thuiller: *Predicting species distribution: offering more than simple habitat models (vol 8, pg 993, 2005).* Ecology Letters, 10(5):435, 2007.

[21] Guisan, A. und N. E. Zimmermann: *Predictive habitat distribution models in ecology.* Ecological Modelling, 135:147–186, 2000.

[22] Hanley, J. A. und B. J. McNeil: *The Meaning and Use of the Area under a Receiver Operating Characteristc (ROC) Curve.* Radiology, 143(1):29–36, 1982.

[23] Hanley, J. A. und B. J. McNeil: *A Method o Comparing the Areas under Receiver Operating Characteristc Curves Derived from the Same Cases.* Radiology, 148(3):839–843, 1983.

[24] Hernandez, P. A., C. H. Graham, L. L. Master und D. L. Albert: *The effect of sample size and species characteristics on performance of different species distribution modeling methods.* Ecography, 29(5):773–785, 2006.

[25] Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones und A. Jarvis: *Very high resolution interpolated climate surfaces for global land areas.* International Journal of Climatology, 25:1965–1978, 2005.

[26] Hirzel, A. und A. Guisan: *Which is the optimal sampling strategy for habitat suitability modelling.* Ecological Modelling, 157(2-3):331–341, 2002.

[27] Hirzel, A. H., G. Le Lay, V. Helfer, C. Randin und A. Guisan: *Evaluating the ability of habitat suitability models to predict species presences.* Ecological Modelling, 199(2):142–152, 2006.

[28] Huisman, j., h. olff und l. fresco: *A hierarchical set of models for species response analysis.* Journal of Vegetation Science, 4(1):37–46, 1993.

[29] Hutchinson, G. E.: *Population Studies - Animal Ecology And Demography - Concluding Remarks.* Cold Spring Harbor Symposia On Quantitative Biology, 22:415–427, 1957.

[30] Jaynes, E. T.: *Information Theory an Statistical Mechanics.* The Physical Review, 106(4):620–630, 1957.

[31] Jooss, R.: *Planungorientierte Einsatz von Habitatmodelle im Landschaftsmaßstab: Kommunale Schutzverantwortung für Zierarten der Fauna*. In: *Treffpunkt Biologische Vielfalt*, Nr. 4. Bundesamt für Naturschutz, Bonn, 2004.

[32] Kölling, C.: *Klimahüllen für 27 Waldbaumarten*. AFZ-Der Wald, 23:1242–1245, 2007.

[33] Kölling, C., L. Zimmermann und H. Walentowski: *Klimawandel: Was geschieht mit Buche und Fichte?*. AFZ-Der Wald, 11:584–588, 2007.

[34] Levins, R.: *Strategy Of Model Building In Population Biology*. American Scientist, 54(4):421–&, 1966.

[35] Levins, R.: *Formal Analysis And The Fluidity Of Science - A Response*. Quarterly Review Of Biology, 68(4):547–555, 1993.

[36] Loiselle, B. A., P. M. Jorgensen, T. Consiglio, I. Jimenez, J. G. Blake, L. G. Lohmann und O. M. Montiel: *Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes?*. Journal Of Biogeography, 35(1):105–116, 2008.

[37] McCune, B.: *Nonparametric Multiplicative Regression for Habitat Modeling*. PDF on Website, 2008.

[38] McKay, R.: *Variants of genetic programming for species distribution modelling - fitness sharing, partial functions, population evaluation*. Ecological Modelling, 146(1-3, Sp. Iss. SI):231–241, 2001.

[39] Nagelkerke, N.: *A note on a general definition of the coefficient of determination*. Biometrika, 78(3):691–692, 1991.

[40] Odenbaugh, J.: *Complex systems, trade-offs, and theoretical population biology: Richard Levin's "strategy of model building in population biology" revisited*. Philosophy Of Science, 70(5):1496–1507, 2003.

[41] Odenbaugh, J.: *The strategy of "The strategy of model building in population biology"*. Biology & Philosophy, 21(5):607–621, 2006.

[42] Oksanen, J. und P. Minchin: *Continuum theory revisited: what shape are species responses along ecological gradients?*. Ecological Modelling, 157(2-3):119–129, 2002.

[43] Ortega-Huerta, M. A. und A. T. Peterson: *Modeling ecological niches and predicting geographic distributions: a test of six presence-only methods*. Revista Mexicana De Biodiversidad, 79(1):205–216, 2008.

[44] Orzack, S. H.: *Discussion: What, If Anything, Is "The Strategy of Model Building in Population Biology?" A Comment on Levins (1966) and Odenbaugh (2003)*. Philosophy of Science, 72(3):479–485, 2005.

[45] Orzack, S. H. und E. Sober: *A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966)*. The Quarterly Review of Biology, 68(4):533–546, 1993.

[46] Peterson, A. T., M. Papes und M. Eaton: *Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent*. Ecography, 30(4):550–560, 2007.

[47] Phillips, S. J., R. P. Anderson und R. E. Schapire: *Maximum entropy modeling of species geographic distributions*. Ecological Modelling, 190:231–259, 2006.

[48] Phillips, S. J. und M. Dudik: *Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation*. Ecography, 31:161–175, 2008.

[49] Phillips, S. J., M. Dudik und R. E. Schapire: *A Maximum Entropy Approach to Species Distribution Modeling*. In: *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[50] Saarchi, S., W. Buermann, H. ter Steege, S. Mori und T. B. Smith: *Modeling distribution of Amazonian tree species an diversity using remote sensing measurements*. Remote Sensing of Environment, 112:2000–2017, 2008.

[51] Schröder, B.: *Habitatmodelle für ein modernes Natruschutzmanagment*. In: Gnauk, A. (Hrsg.): *Theorie und Modellierung von Ökosystemen - Workshop Köpingsee 2000*, S. 201–224, 2000.

[52] Senyk, N. A. und J. Sanchez: *Development of a dynamic optimal habitat model to describe the spatial and temporal habitat distributions of giant kelp, Macrocystis pyrifera*. CEQA Final Report, University of California, Santa Barbara, 2005. Veröffentlicht im eScholarship Repository, University of California. http://repositories.cdlib.org/ucmarine/ceqi/001.

[53] Shan, Y., R. McKay und D. Paull: *Building ecological models using genetic programming*. verfügbar über: http://citeseer.ist.psu.edu/545087.html.

[54] Stockwell, D. und D. Peters: *The GARP modelling system: problems and solutions to automated spatial prediction*. International Journal Of Geographical Information Science, 13(2):143–158, 1999.

[55] Tarkesh, M.: *Effect of sampling design on predictive vegetation mapping and community-response curve*. unveröffentlicht, eingereicht, 2008.

Anlagen

Teil I - verwendete Software

ARCVIEW 3.0 Zur Stichprobenerstellung verwendet.

DevC++ Bloodshed C++ - IDE. Programmierung des HabitatGP.

gcc 3.4.2 GNU C++-Compiler mit Dev C++ mitgeliefert

GIMP 2.4.0 Bearbeitung und Konvertierung von Grafiken

ILWIS 3.0 Academic Freie GIS-Software. Zur Handhabung der PNV-, Umwelt- und Habitatkarten und zur erstellung der Abbildungen für die Arbeit verwendet. Verfügbar unter http://52north.org/

Mathematica 5.0 Auswertung der vom HabitatGP gelernten Funktionen.

MaxEnt 3.2.1 Berechnung der Maximum-Entropie-Habitatmodelle

ModelEvaluator Selbstgeschriebenes Programm zu Berechnung von ROC-Kurven, AUC und PE-Ratio-Graphen.

OpenOffice 2.4.0 OOCalc Berechnung von R^2 aus den Ausgabedateien von MaxEnt, Erstellung der Stichprobendateien (im CSV-Format) als Input für MaxEnt und HabitatGP. OODraw, Erstellung einiger Grafiken.

SPSS Version 15 statistischen Auswertungen

TeXnicCenter 1 Beta 7 Erstellung des Latexdokuments und Generierung des PDF.

Teil II - Abbildungen der Modelle aus Kapitel 4.1.2

Abbildung II.1: Stichprobe F50-F59

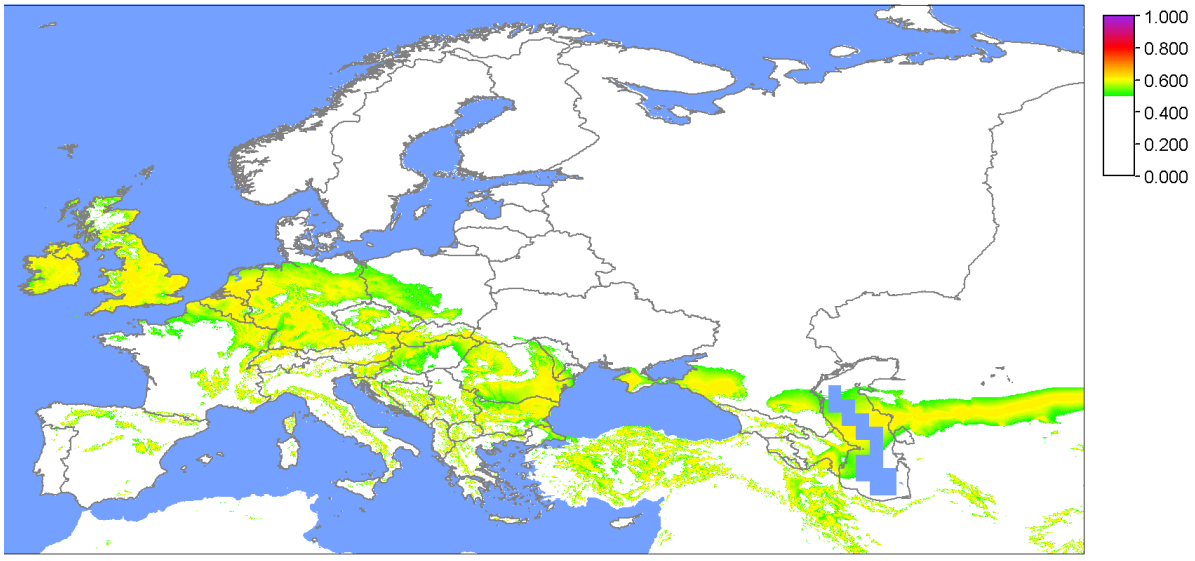


Abbildung II.2: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D1

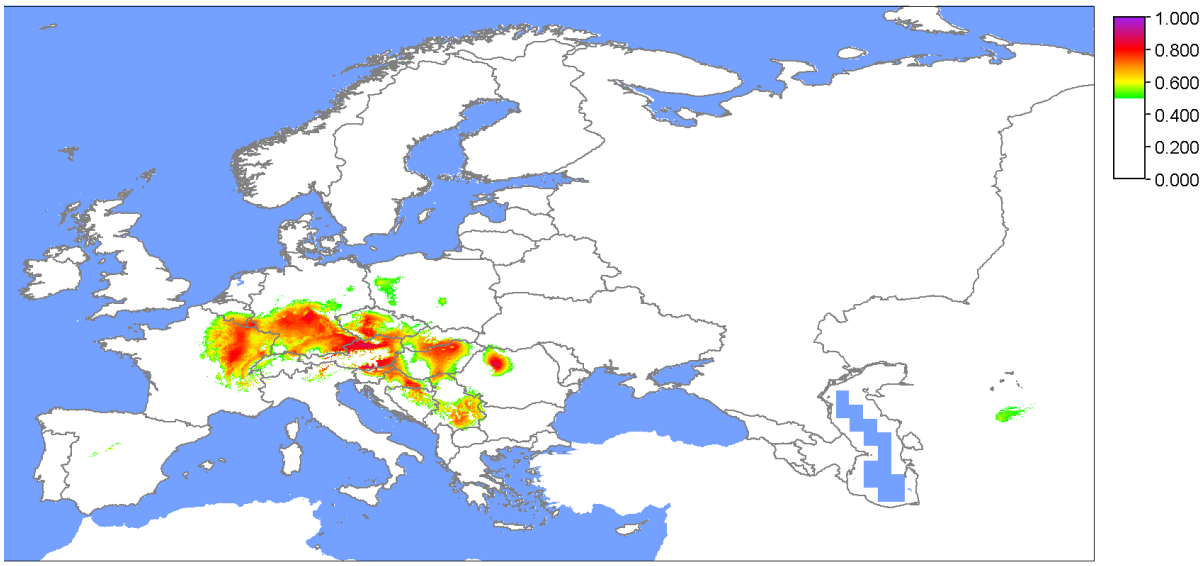


Abbildung II.3: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D2

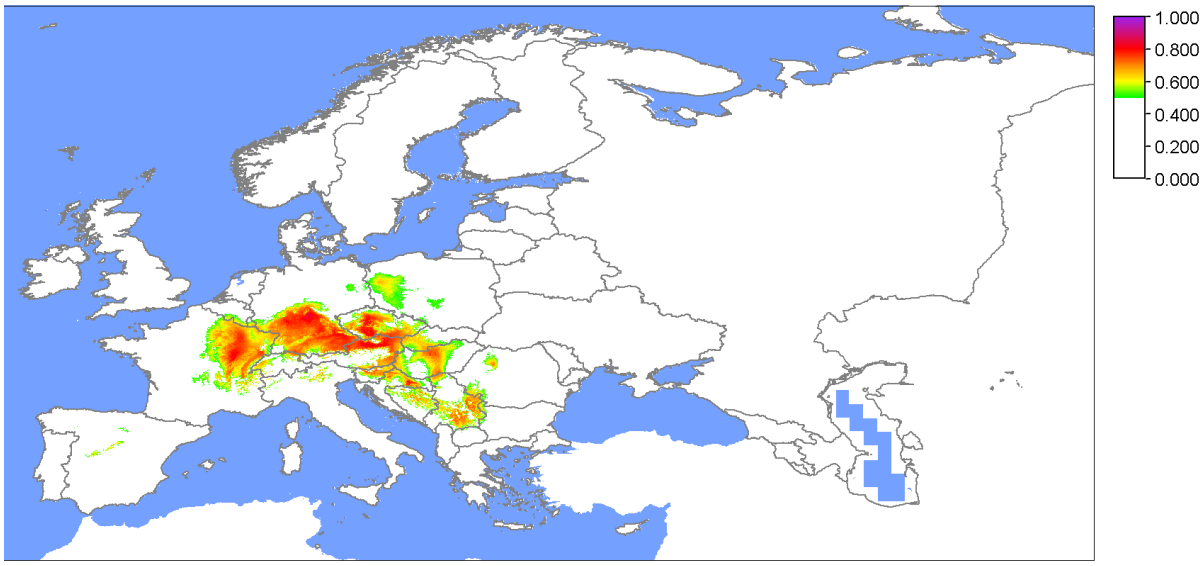


Abbildung II.4: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D3

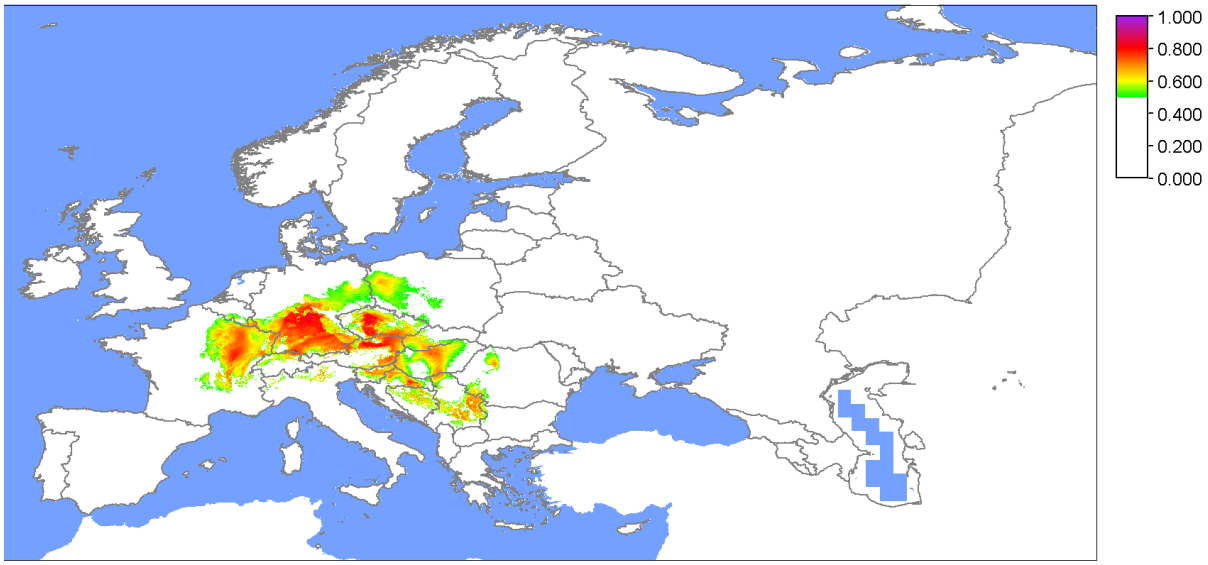


Abbildung II.5: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D4

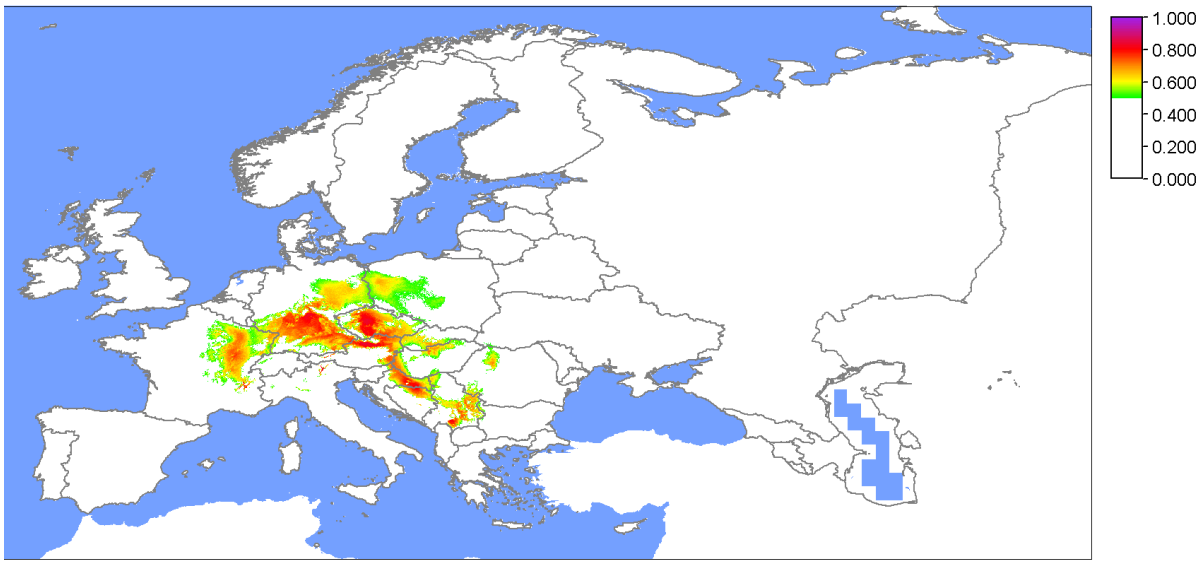


Abbildung II.6: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D5

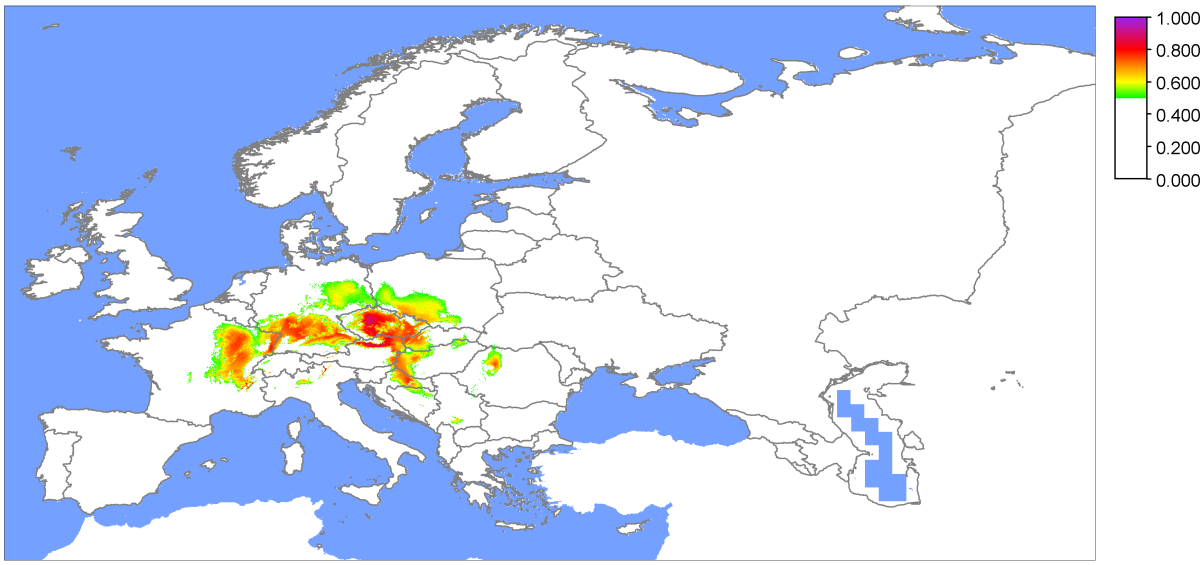


Abbildung II.7: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D6

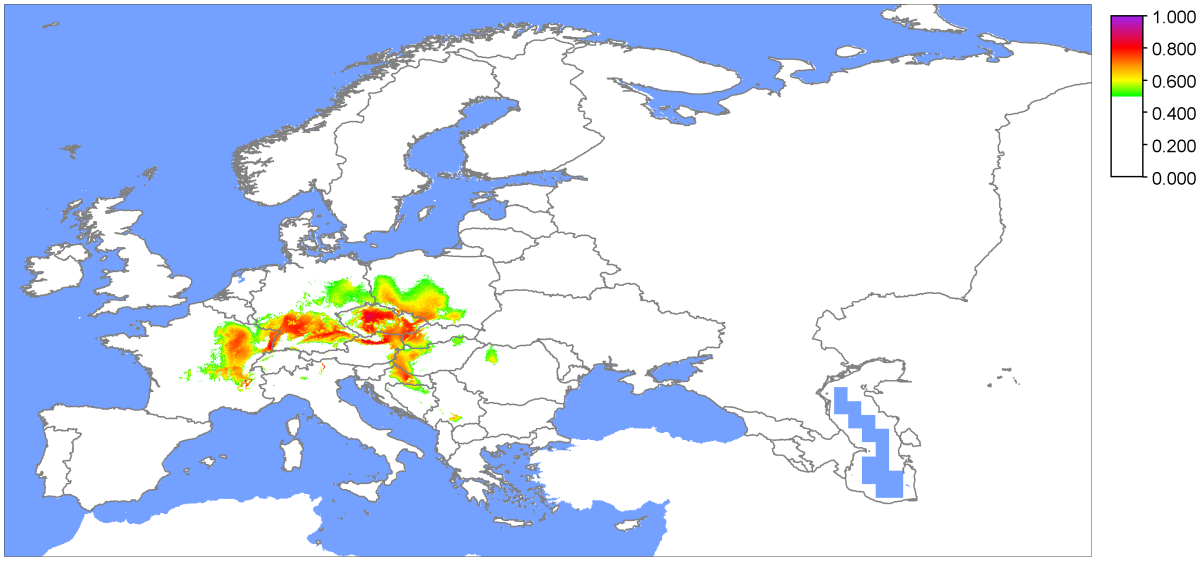


Abbildung II.8: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D7

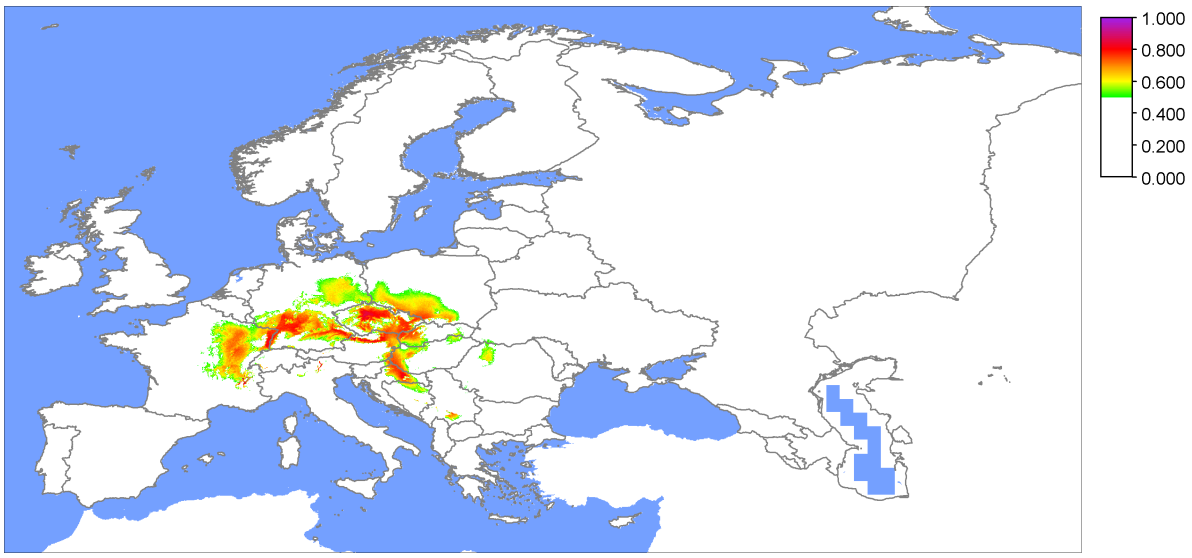


Abbildung II.9: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D8

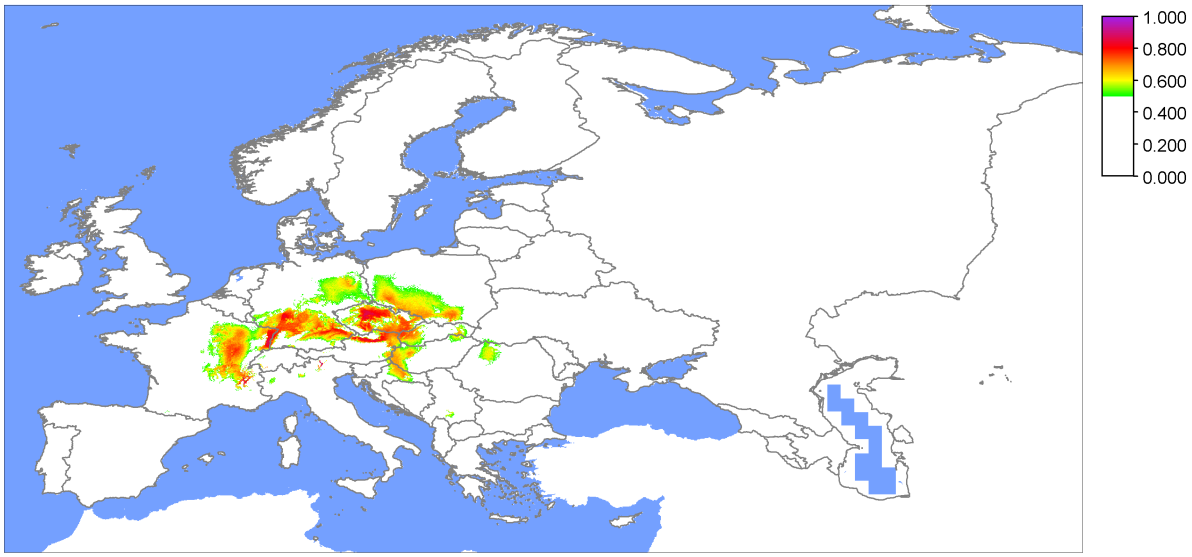


Abbildung II.10: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D9

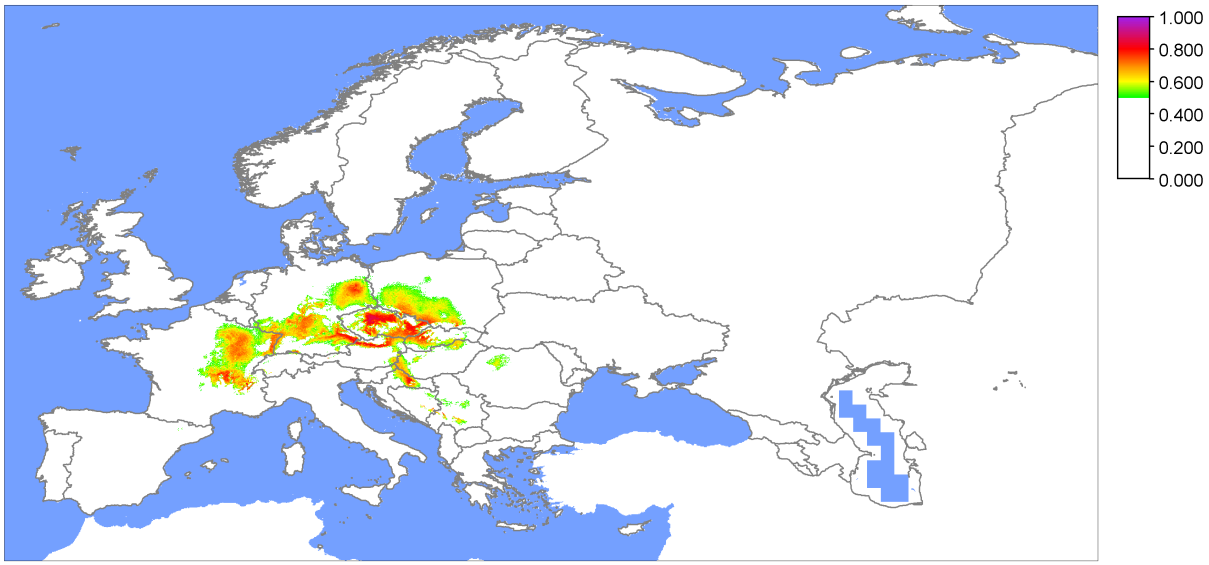


Abbildung II.11: Mit MaxEnt berechnete logistische Wahrscheinlichkeitsverteilung der Eichen-Hainbuchen-Wälder(F50-F59). Modell D10

Teil III - Flussdiagramm des HabitatGP

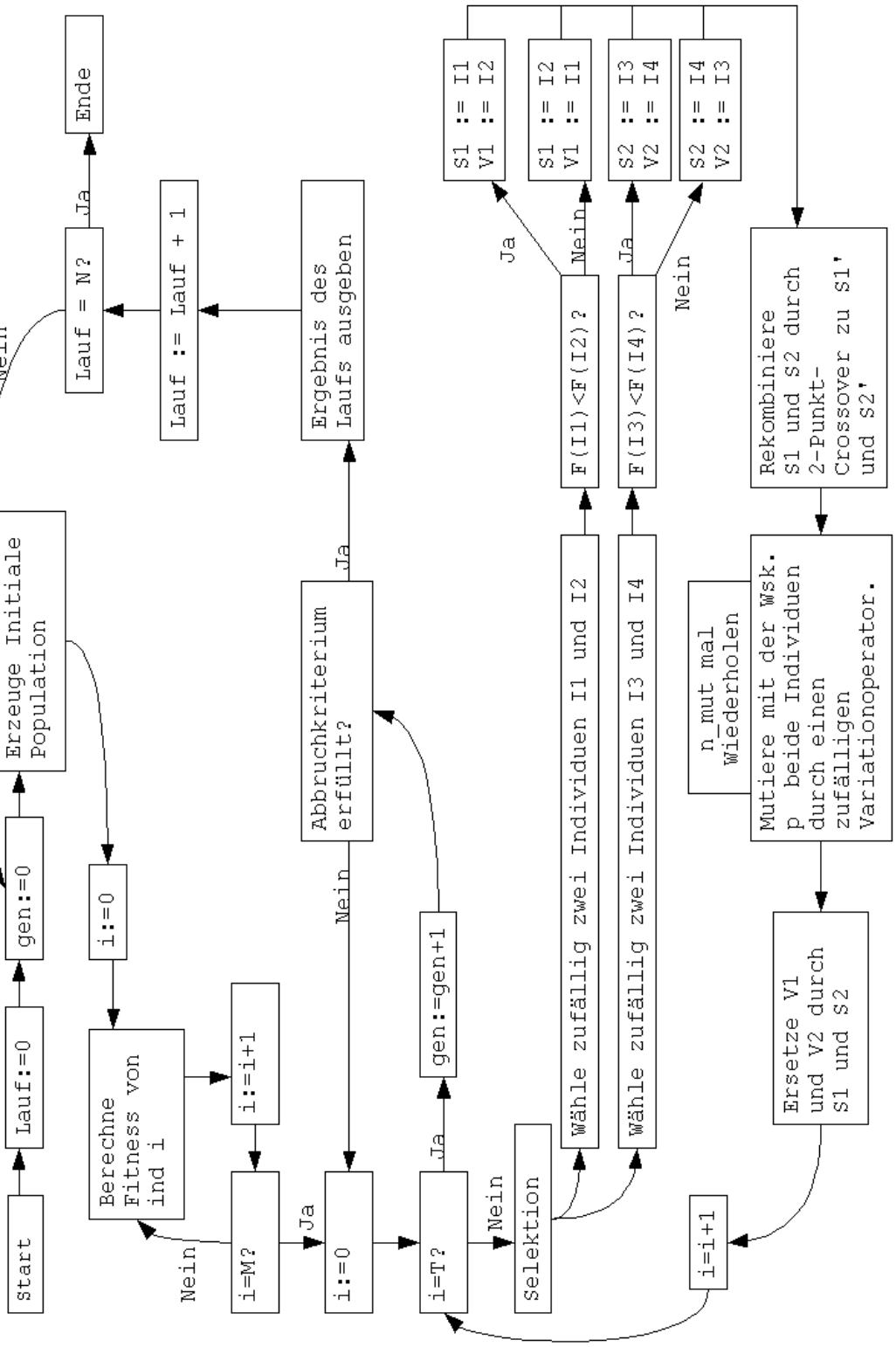


Abbildung III.1: Flussdiagramm des Ablaufs des HabitatGP

Teil IV - Lernverläufe des HabitatGP

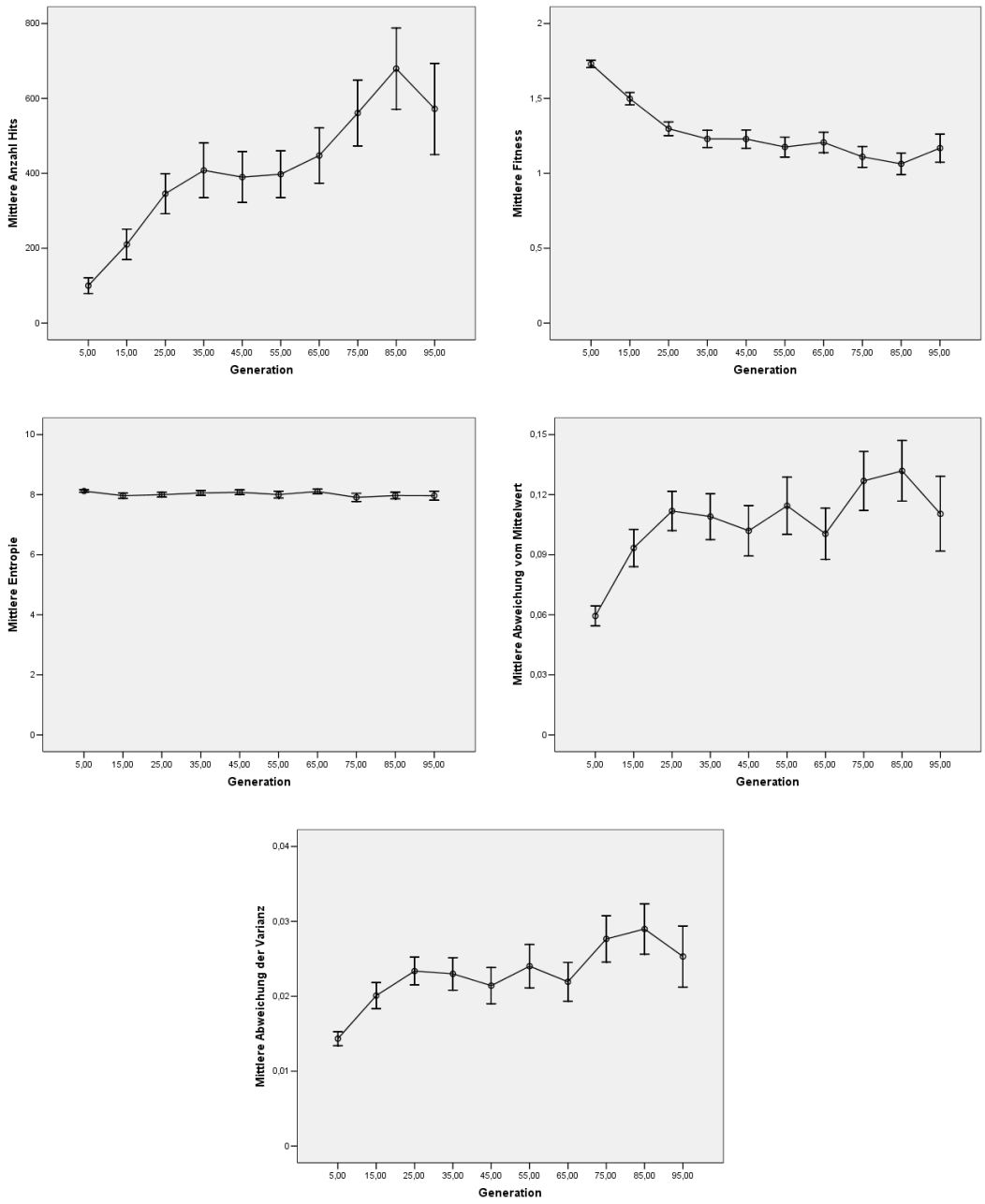


Abbildung IV.1: Mittelwerte jeder Komponente der Zielfunktion Z_{FH} für T1. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

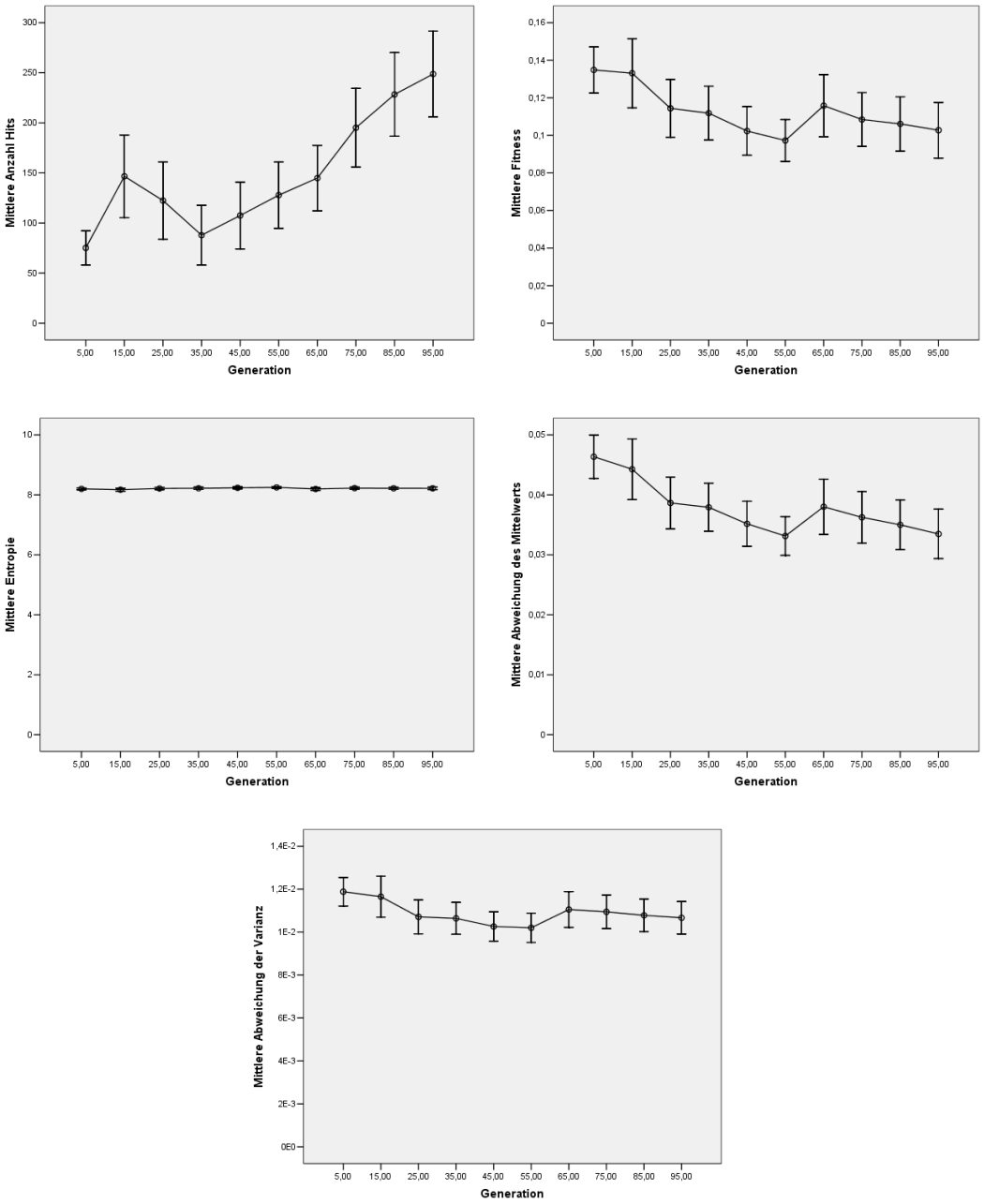


Abbildung IV.2: Mittelwerte jeder Komponente der Zielfunktion Z_{EntExp} für T1. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

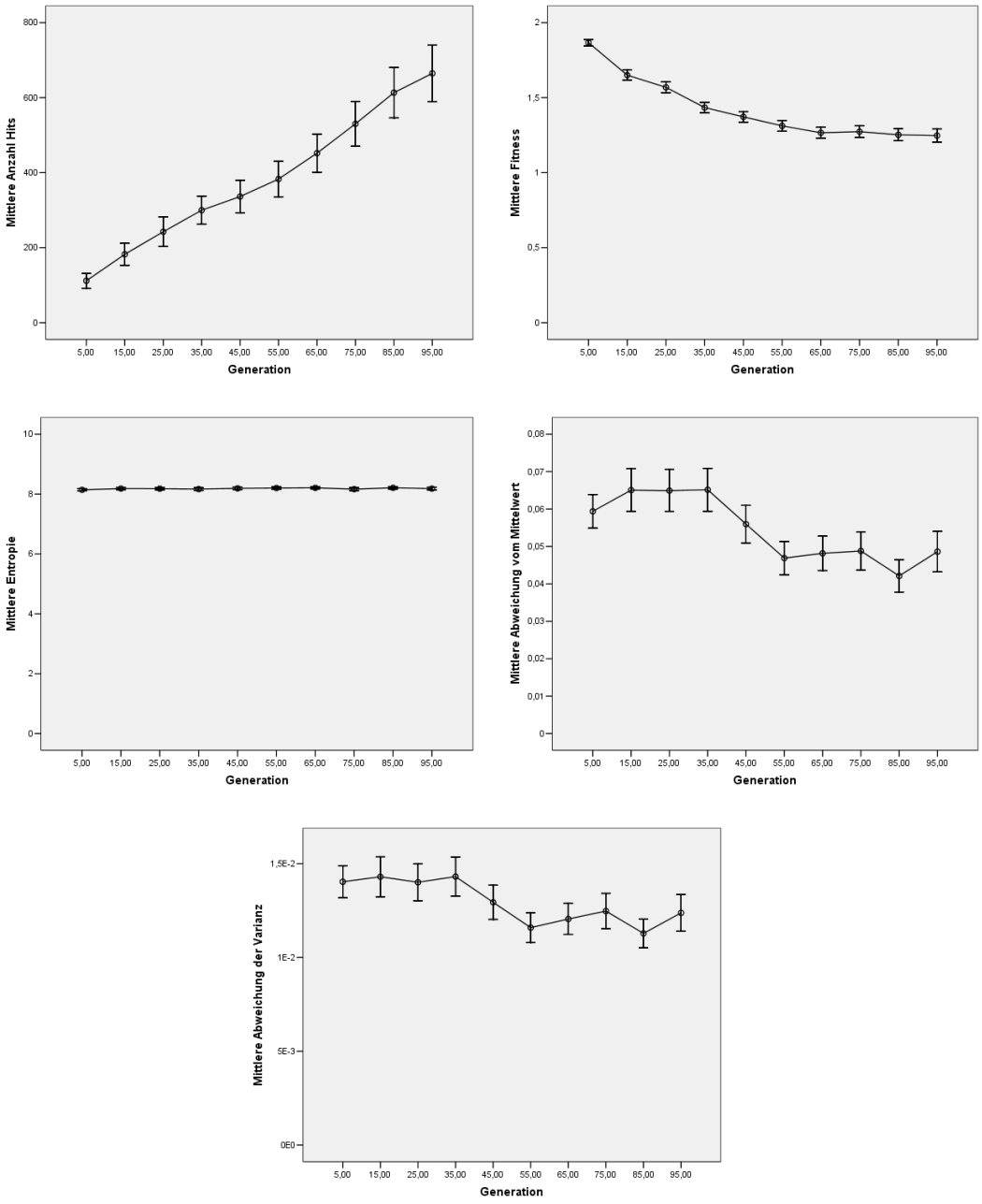


Abbildung IV.3: Mittelwerte jeder Komponente der Zielfunktion Z_{Full} für T1. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

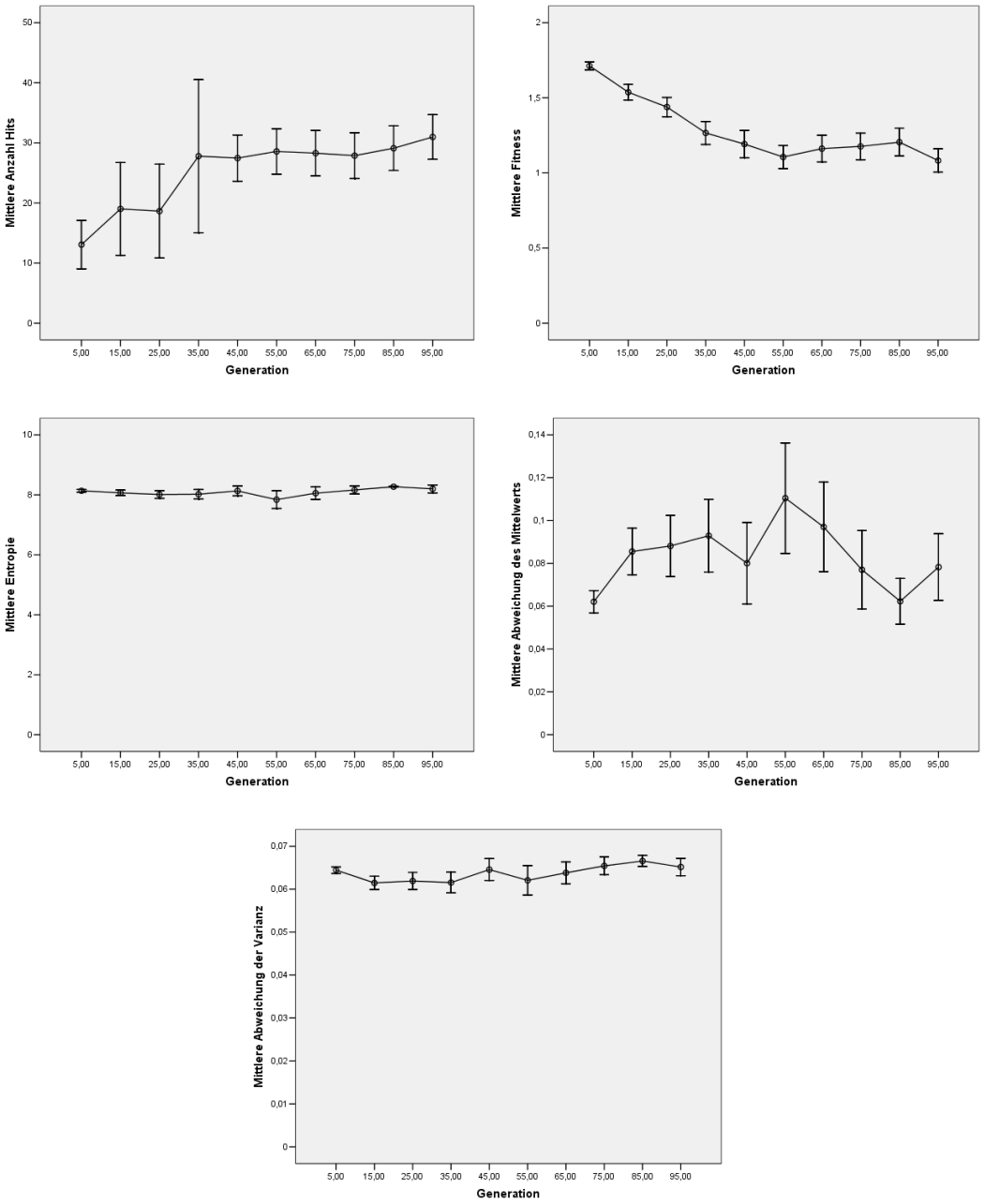


Abbildung IV.4: Mittelwerte jeder Komponente der Zielfunktion Z_{FH} für T2. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

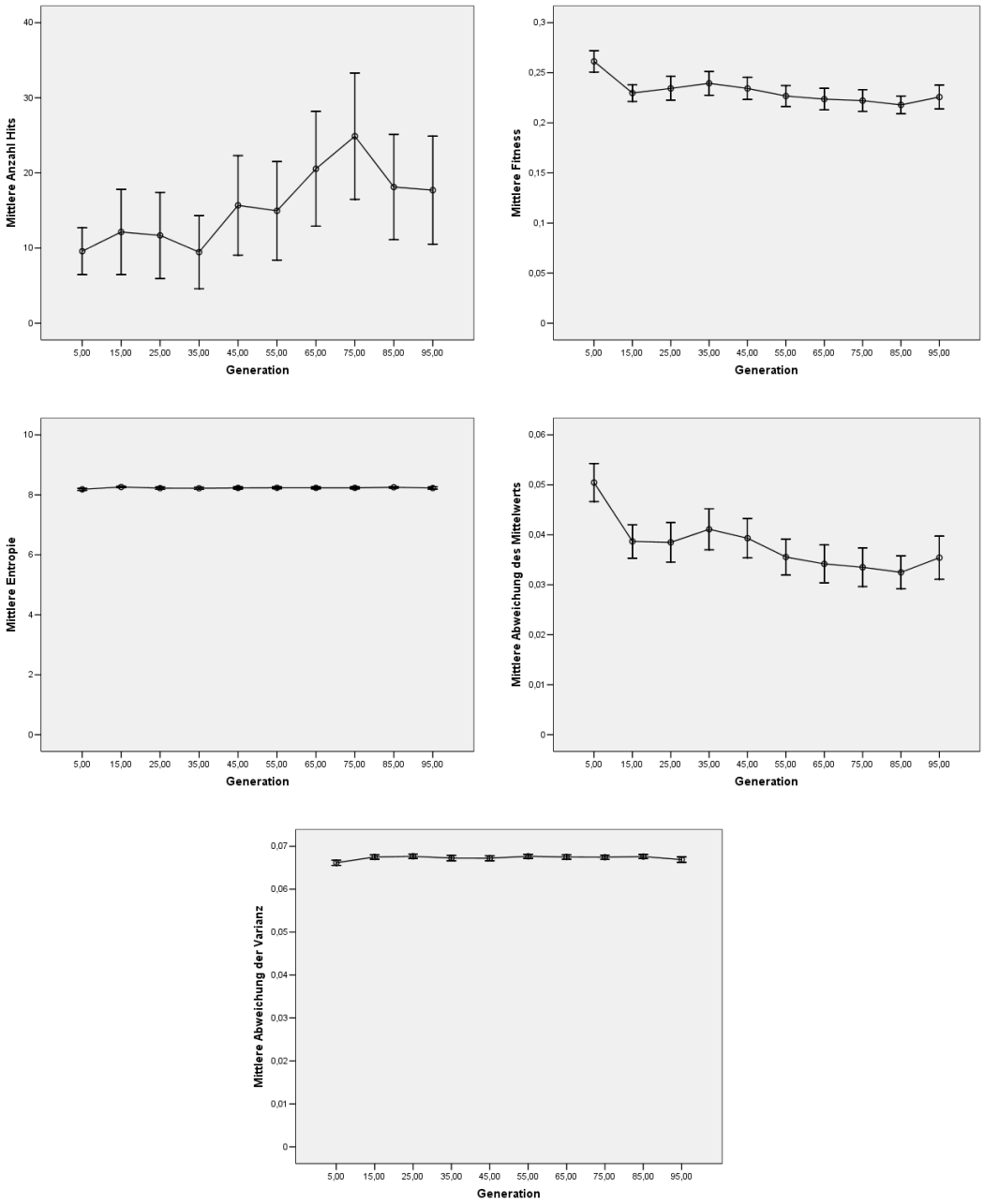


Abbildung IV.5: Mittelwerte jeder Komponente der Zielfunktion Z_{EntExp} für T2. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

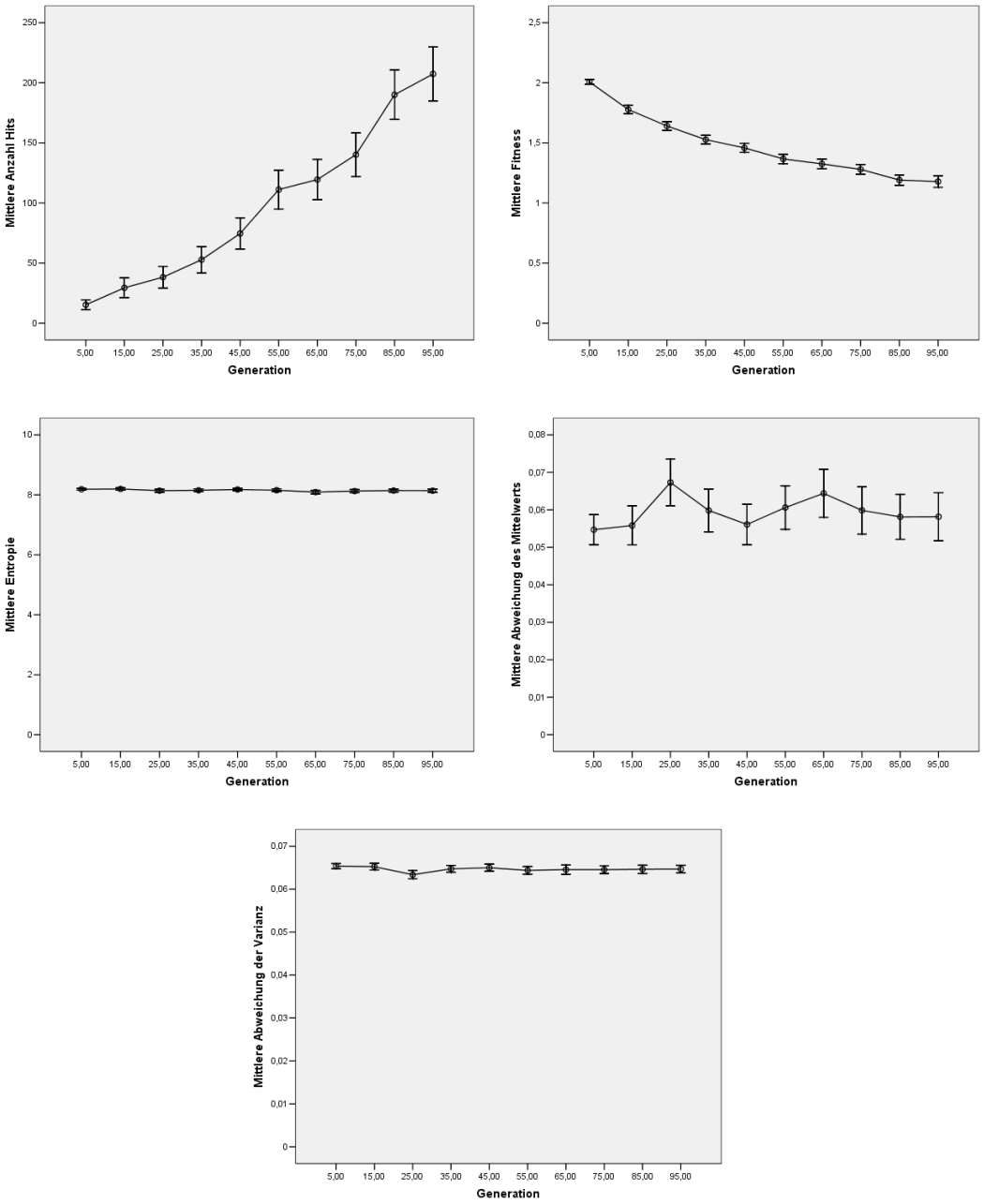


Abbildung IV.6: Mittelwerte jeder Komponente der Zielfunktion Z_{Full} für T2. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

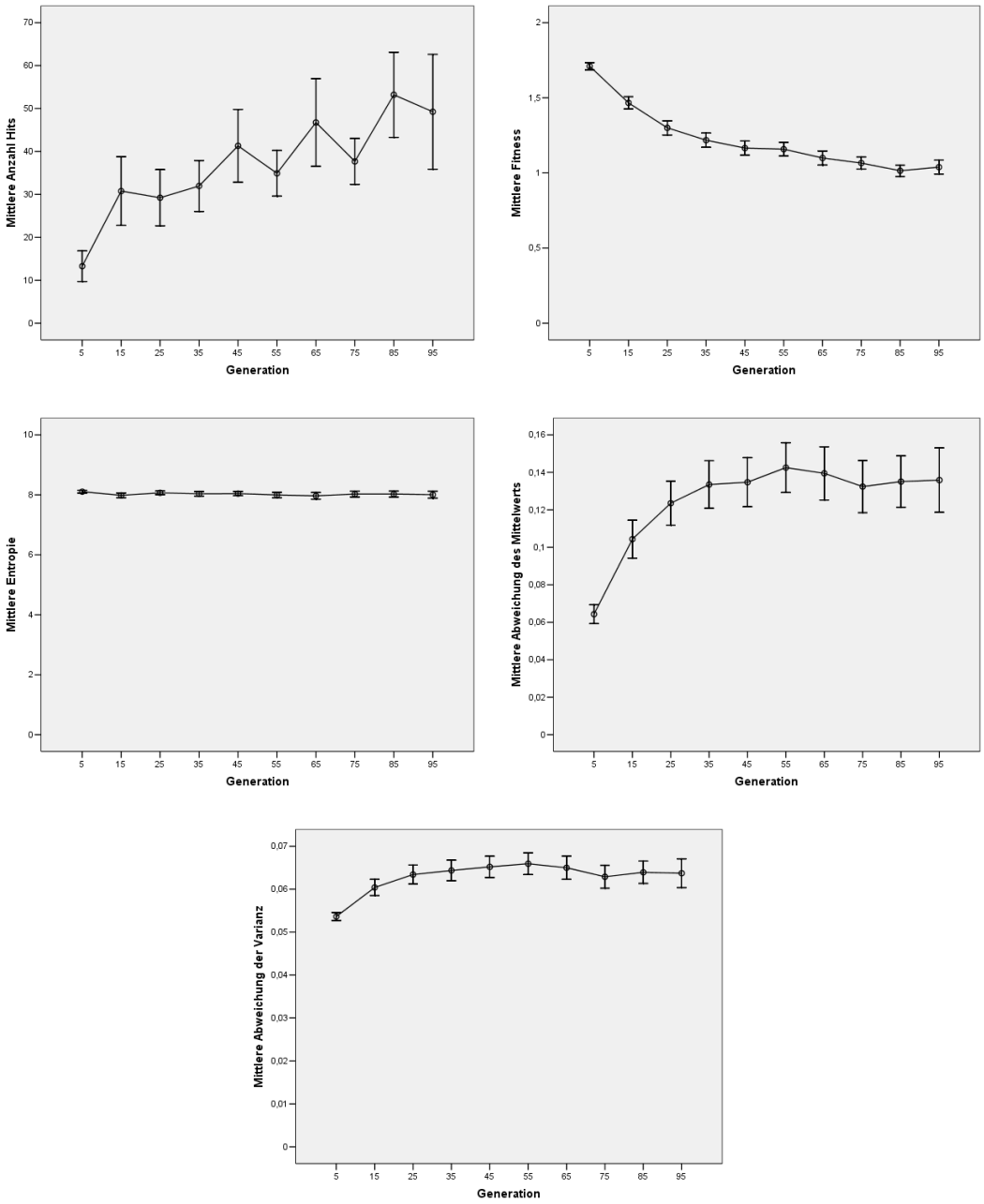


Abbildung IV.7: Mittelwerte jeder Komponente der Zielfunktion Z_{FH} für T3. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

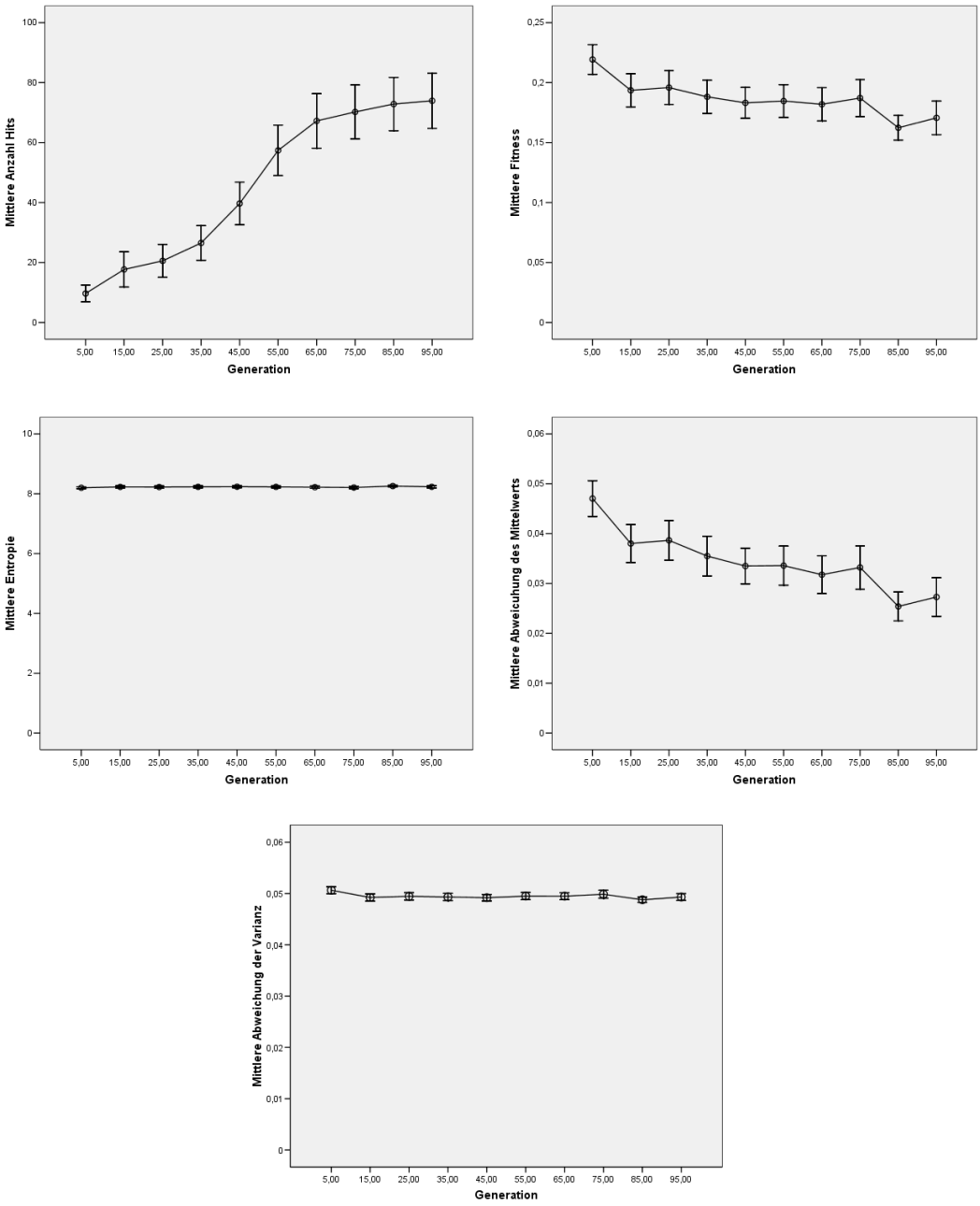


Abbildung IV.8: Mittelwerte jeder Komponente der Zielfunktion Z_{EntExp} für T3. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

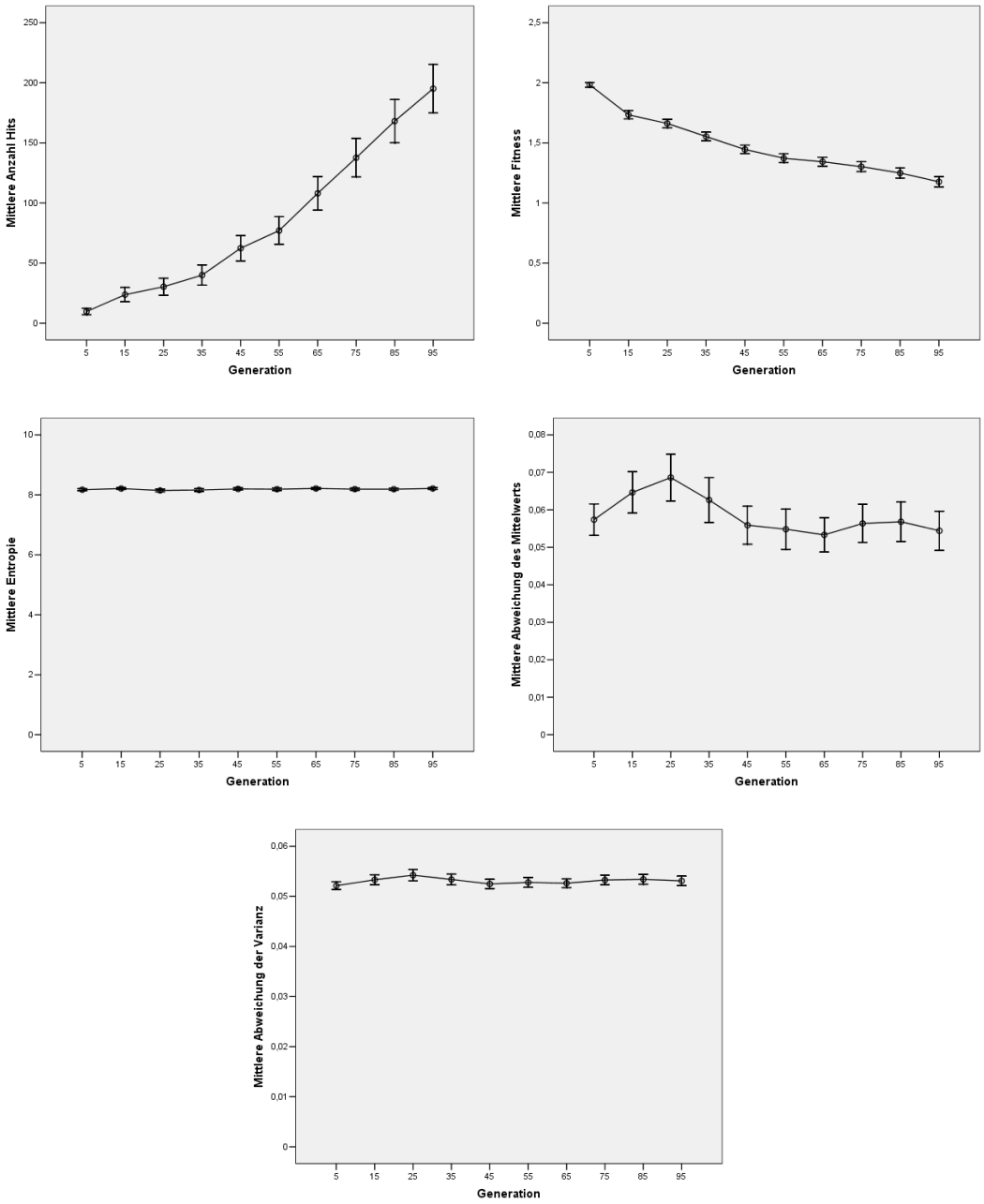


Abbildung IV.9: Mittelwerte jeder Komponente der Zielfunktion Z_{Full} für T3. Die Fehlerbalken zeigen das 95%-Konfidenzintervall.

Teil V - DVD-Inhalt

-Diplomarbeit.pdf - Diese Arbeit als PDF-Dokument

/Daten

Alle Stichprobendateien im CSV-Format
- testsampling1.csv
- testsampling2.csv
- testsampling3.csv
- trainF34-F69.csv
- trainF50-F54.csv
- trainF55-F57.csv
- trainF58-F59.csv
- trainF50-F59.csv

./Umwelt Alle Umweltdaten
- artificialmap.asc
- deutschland_bio1.asc /bio2/bio3
- europe_tmin<monat>.asc - <monat> ist eine Zahl zwischen 1 und 12
- europe_tmax<monat>.asc - <monat> ist eine Zahl zwischen 1 und 12
- europe_prec<monat>.asc - <monat> ist eine Zahl zwischen 1 und 12
- europe_bio<nummer>.asc - <nummer> ist eine Zahl zwischen 1 und 19

./Umwelt 2080 analog zu /Umwelt

/Modelle

./MaxEnt
- G1, G2, G3
- K0_1-K0_10, K+_1-K+_10, K-_1-K-_10
- D1 - D10
- F34-F69
- F50-F59
- F50-F54
- F55-F57
- F58-F59

./HabitatGP
- T1, T2, T3 jeweils nach Zielfunktion aufgeschlüsselt
enthalten die Dateien zu den 10 Replikaten.

/Auswertung

SPSS-Dateien

- AUCR2.sav - Enthält AUC_train,AUC_test und R^2 für alle Modelle aus 4.2
- PercentageContribution.sav - Enthält den prozentualen Anteil jeder Variable für die Modelle aus 4.2
- GP_Testprobleme.sav - Beste Fitness und Laufzeiten
- <Testproblem>_<Zielfunktion>_Fitness_AlleReplikate.sav - Datenbasis für die Diagramme aus Teil IV der Anlagen.

/Programme

- MAXENT v.3.2.1
- HabitatGP
- ModelEvaluator: Selbst erstelltes Programm zur Berechnung der PE-Graphen

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Jena, den 31. Juli 2008. . . . . .
Dennis Görlich